*Ю.Н.Толстова*

# АНАЛИЗ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

**Методология, дескриптивная статистика, изучение связей между номинальными признаками**

**Рекомендовано Министерством образования Российской Федерации в качестве учебного пособия для студентов кафедр и факультетов социологии университетов России**

**Москва**

**Научный мир**

**2000**

**УДК 519.2 : 316**
**Т 53**
**ББК 60.56;60.6**
**ISBN 5-89176-086-Х10**

**Ю.Н.Толстова**
**АНАЛИЗ СОЦИОЛОГИЧЕСКИХ ДАННЫХ**
**Методология, дескриптивная статистика, изучение связей между номинальными признаками.** –М.: Научный мир, 2000.- 352с.

Учебное пособие отвечает курсу "Анализ социологических данных", читаемому автором студентам-социологам нескольких вузов г.Москвы. В нем рассматривается ряд методологических положений, отличающих эту дисциплину – анализируется класс соответствующих социологических задач, прослеживается связь с математической статистикой, раскрывается специфика применения алгоритмов анализа данных именно в социологии. Большинство положений используется в процессе подробного рассмотрения ряда конкретных методов. Отобраны методы, наиболее адекватные потребностям социологии – традиционные методы описательной статистики и методы анализа связей между номинальными признаками. Многие рассматриваемые алгоритмы изучения связей слабо отражены в отечественной литературе. Предлагается классификация подходов к анализу связей, отвечающая естественной логике социолога-эмпирика.

Книга рассчитана на студентов- и аспирантов-социологов, на всех лиц, желающих эффективно изучать социологическую эмпирическую информацию. Предполагается знание курсов по общей социологии, методике социологических исследований, теории вероятностей и математической статистике, теории измерений в рамках обычных вузовских программ.

**Публикуется при финансовой поддержке Международного фонда RSS, Contract No.: 854/1997 и (частично) Российского фонда фундаментальных исследований, проект № 99-06-80065**



**ISBN 5-89176-086-Х10**

# Содержание

## *Часть 2.*

## ОПИСАТЕЛЬНАЯ СТАТИСТИКА. ИЗУЧЕНИЕ СВЯЗИ МЕЖДУ НОМИНАЛЬНЫМИ ПРИЗНАКАМИ

## ВВЕДЕНИЕ

Настоящая работа является **учебным пособием**, отвечающим курсу “Анализ социологических данных”, читаемому автором для студентов социологических факультетов Московского государственного университета им. М.В.Ломоносова и ряда других вузов Москвы (программа курса была опубликована [Толстова, 1994, 1996а]). Книга состоит из двух частей. В первой рассматриваются методологические аспекты процесса анализа данных в социологии. Вторая посвящена описанию отдельных методов. Поясним, почему возникла потребность использования такой структуры текста.

В наше время каждый социолог понимает, что собранные им данные так или иначе надо "анализировать" (конечно, с помощью математических методов). Практически в каждом учебном заведении, готовящем социологов, предусматривается преподавание предмета, название которого фигурирует в заголовке настоящей книги. Но, на наш взгляд, далеко не всегда совокупность действий, называемая анализом социологических данных, понимается правильно. В первую очередь, мы имеем в виду то, что эта совокупность действий не всегда трактуется как некоторый специфичный процесс, не сводящийся ни к какому набору математических приемов и органично вписывающийся в содержательную ткань социологического исследования. Непонимание же сути указанного процесса, по нашему мнению, приводит к неэффективному использованию математического аппарата, и, более того, к получению выводов, противоречащих реальности. Неадекватное отношение к процессу анализа данных не является случайным.

Несмотря на то, что в литературе имеется довольно много отдельных публикаций, посвященных изучению специфики процесса анализа данных в социологии, существование научной ветви с названием "анализ данных социологического исследования", или "анализ социологических данных" пока наукой не "узаконено". И, вероятно, разумно полагать, что такое положение дел сохранится до тех пор, пока не будет создан и признан научной общественностью какой-либо учебник по дисциплине с указанным наименованием. Подобный учебник должен раскрывать соответствующие приемы и методы как нечто специфичное именно для социологии. Такого учебника пока нет не только у нас в стране, но и за рубежом (на Западе имеется огромное число книг, в которых так или иначе фигурирует словосочетание "анализ данных"; но в этих книгах, по нашему мнению, не достаточно полно и глубоко рассматривается проблема "стыковки" рассматриваемых математических методов именно с социологией).

Важно также отметить, что некоторые обстоятельства иногда заставляют сомневаться и в существовании дисциплины, именуемой просто "анализ данных". Свидетельством этого

можно считать, например, то, что упомянутый термин в литературе понимается по-разному (см., например, [Толстова, 1995а]). Этот факт тоже существен для практики: чтобы получать корректные выводы, мы должны четко понимать, когда, в каких именно условиях и с какой целью можно использовать анализ данных, а это немыслимо без ясного представления о том, что это такое. Ответу на соответствующий вопрос и посвящена первая часть работы. Она отвечает нескольким первым лекциям курса, читающегося автором. Многие из рассмотренных в ней положений конкретизируются при рассмотрении реальных методов анализа данных во второй части книги. Перейдем к более подробному описанию каждой из частей.

В **первой части** книги разъясняется, что означает словосочетание "анализ социологических данных", каков смысл каждой из его составляющих. Хотелось бы, чтобы в результате у читателя сформировалось четкое представление о том, с какой областью науки мы имеем дело, каково место этой области в общей структуре человеческого знания о мире и, главное, зачем все нижеизложенное нужно социологу в его практической работе. Можно сказать, что в первой части речь идет о той "среде", в которой должен действовать каждый социолог, пытающийся "выудить" какие-либо закономерности из "моря" полученной им эмпирической информации.

Основные наши рассмотрения сводятся к демонстрации сути статистических закономерностей, на выявление которых нацелен анализ данных; к проведению границы между анализом данных и математической статистикой, которая тоже предназначена для поиска статистических закономерностей; к рассмотрению некоторых аспектов анализа данных, специфичных именно для социологии.

Отметим, что поначалу мы будем использовать термин "анализ данных", понимая соответствующую область знания интуитивно, как нечто рядоположенное с такой ветвью науки, как "математическая статистика". Далее определим понятие "анализ данных" более строго, четко выявив границы его размежевания с математической статистикой (раздел 4). Но предварительно нам потребуется рассмотреть подробнее понятие статистической закономерности и проанализировать его значение для социолога (раздел 1); показать, что социолог не может в своей работе обойтись без математики (раздел 2); продемонстрировать, что при поиске статистических закономерностей естественно использовать именно ту ветвь математики, которая называется "математическая статистика" (раздел 3). Развивая далее соответствующие положения, мы сможем в рамках анализа данных вычленить тот его фрагмент, который можно связать с решением именно социологических задач (раздел 5) (хотя, конечно, мы не можем полностью “отречься” от социологии и в первых четырех разделах).

Несколько слов следует сказать о приведенных в конце первой части Примечаниях. Дело в том, что некоторые из них носят принципиальный характер, касаются вопросов, актуальных для современной социологии, но пока не решенных до конца (речь идет в основном о методологических проблемах получения социологического знания). Сочтя неуместным вставлять соответствующие рассуждения в основной текст, посвященный сравнительно узкой проблематике, мы позволили себе привести их в сносках, сделав последние иногда довольно пространными. Хотелось бы, чтобы читатель (особенно студент-социолог) задумался относительно затронутых в Примечаниях вопросов.

**Вторая часть** содержит описание конкретных методов анализа данных и делится на два относительно автономных раздела:

- изложение методов т.н. описательной (дескриптивной) статистики - выборочного представления одномерного вероятностного распределения и расчета его основных параметров (мер средней тенденции и показателей разброса);
- описание простейших методов изучения связей между номинальными признаками

Конечно, нельзя считать, что этими методами должен ограничиваться круг знаний социолога в области анализа данных. Так, на практике может возникнуть потребность изучения связей между признаками, значения которых получены по шкалам более высокого типа, чем номинальные. Однако мы сознательно ограничились лишь номинальным уровнем измерения: номинальные данные чаще используются в социологии и являются более надежными. Кроме того, методы, рассчитанные на работу со шкалами более высокого типа, обычно изучаются студентами-социологами в курсе математической статистики (имеются в виду, например, коэффициенты связи для ранговых признаков, элементы дисперсионного и факторного анализа).

Часто в практической работе социолога требуется использование более сложных методов - например, логлинейного или причинного анализа. Они здесь тоже не рассматриваются.

Представляется также, что, помимо методов расчета показателей дескриптивной статистики и изучения связи между переменными можно выделить по крайней мере еще два мощных класса методов, отвечающих задачам, встающим при анализе данных практически в каждом эмпирическом социологическом исследовании: методы классификации и методы поиска латентных переменных [Толстова, 1994]. В данной работе мы их рассматривать не будем и говорим о них только для того, чтобы более четко оттенить значимость для социологии именно тех подходов, которые рассматриваются в настоящей книге.

Почти все представленные во второй части методы известны, описаны в литературе. Поэтому, вероятно, требуется пояснить, почему мы решились включить их в книгу, почему их

описание представляется нам **актуальным**. Рассмотрим интересующие нас аспекты состояния учебно-методического обеспечения социологического образования.

Сначала - об **отечественной литературе**. В течение 70-х - 80-х годов в стране было опубликовано довольно много работ, предназначенных для ознакомления широких кругов социологов с наиболее перспективными для решения социологических задач математическими методами (см., например, [Паниотто, Максименко, 1982], серию коллективных монографий, выпущенных Институтом социологии СССР [Интерпретация и анализ ..., 1987; Математический анализ и ..., 1989; Статистические методы ..., 1979; Типология и классификация ..., 1982], переведенную с английского языка книгу [Гласс, Стэнли, 1976]). Однако положение дела нельзя считать удовлетворительным. Причин тому несколько.

Во-первых, опубликованные на русском языке работы, содержащие описание рассматриваемых методов (и ориентированные на читателя-социолога, о других мы пока не говорим), в последние годы стали трудно доступными для студентов: книги были изданы давно, нужные страницы в имеющихся в библиотеках экземплярах зачастую утрачены; во многих вузовских библиотеках этих работ нет, поскольку соответствующие социологические подразделения организованы существенно позже выхода книг в свет (названия работ, о которых идет речь, включены в библиографию, приведенную в конце книги).

Во-вторых в нашей литературе нет работ, в которых наиболее актуальные методы интересующего нас плана были бы сведены воедино.

В-третьих, некоторые методы, представляющиеся весьма полезными для социологов, не описаны с достаточной подробностью и четкостью на русском языке с ориентацией на читателя-социолога (это касается, например, методов анализа фрагментов таблицы сопряженности, номинального регрессионного анализа, логлинейного анализа и т.д.). Ряд методов вообще не затрагивается в ориентированной на социолога отечественной литературе (например, линейные обобщенные модели - в частности, логистическая регрессия, пробит-модели, пуассонова регрессия; многие алгоритмы анализа отношений преобладания и т.д.).

В-четвертых, в имеющихся публикациях не учитываются полученные отечественными исследователями в последние годы результаты в области анализа таблиц сопряженности (например, [Ростовцев, 1996, 1997; Витяев, Логвиненко, 1998], а также методические наработки, касающиеся специфики использования математического аппарата именно в социологии (например, мало внимания уделяется анализу моделей, заложенных в математических

алгоритмах, сопряжению этих моделей с содержательными социологическими постановками задач).

В-пятых, не все имеющиеся в отечественной литературе (даже такой, которая ориентирована на социолога) описания интересующих нас методов написаны языком, понятным студентам-гуманитариям (проблема преподнесения таким студентам дисциплин, так или иначе использующим математический аппарат, хорошо известна; соответствующими недостатками, к сожалению, обладают и многие из названных выше работ, что стало видно лишь по мере накопления отечественного педагогического опыта; первые социологические факультеты в российских вузах были организованы в 1989 году).

Несколько слов скажем об известных нам **западных работах**, лежащих в интересующем нас русле. Пальма первенства в разработке многих рассматриваемых в настоящей книге методов принадлежит западным ученым. Методы активно используются на практике, в том числе в эмпирической социологии. Учебно-методическое обеспечение социологического образования на Западе и по качеству, и по количеству несоизмеримо с нашим. Поэтому, конечно, здесь есть что заимствовать.

В западной литературе имеются прекрасные книги, являющиеся по существу адаптированными для читателя-гуманитария учебниками одновременно по теории вероятностей, математической статистике, многомерному статистическому анализу (см., например, [Bluman,1995; Diamantopoulos et al., 1997; Hinton, 1995; Kachigan, 1986; Neter et al., 1990; Sirkin, 1995; Tabachnick et al., 1996; Walsh, 1990]). Мы не раз убеждались в том, что студенты-социологи прекрасно усваивают изложенный в них материал. Эти учебники содержат описание основных свойств распределений одномерных случайных величин, элементы теории статистического оценивания параметров и проверки статистических гипотез, основы регрессионного, дисперсионного, факторного и других видов числового многомерного анализа.

Однако в названных книгах не затрагивается ряд интересующих социолога моментов. Выделим два. Во-первых, математико-статистические подходы не "привязаны" к "нехорошей" социологической ситуации. Так, при описании способов построения гистограмм не анализируются методы работы с пропущенными данными, не рассматривается проблема разбиения диапазона изменения признака на интервалы и т.д. Во-вторых, многие важные для решения социологических задач алгоритмы в названных учебниках просто не рассматриваются. Так, подход к изучению наиболее важного для социолога объекта - частотной таблицы - затрагивается, как правило, лишь в традиционном для математической статистики варианте -

рассматривается способ измерения связи между двумя переменными с помощью критерия Хи-квадрат. В большинстве учебников остаются в стороне многие методы изучения номинальных данных, отражающие наиболее естественную логику рассуждений социолога (например, описанные ниже алгоритмы типа AID). Однако это не означает отсутствие соответствующей методической литературы. Напротив, работ интересующего нас характера много.

Прежде всего отметим книгу [Agresti, 1990], в которой сравнительно простым языком описаны многие подходы, вообще не описанные в отечественной ориентированной на социолога литературе, но давно известные и ставшие классикой на Западе (многие логлинейные, логит-, пробит- модели, ряд моделей логистической регрессии, алгоритмы анализа отношений преобладания и т.д.). Не отражены с достаточной полнотой эти методы и в переводной литературе. Хотя здесь имеет смысл назвать ставшую библиографической редкостью работу [Аптон, 1982], содержащую описание ряда методов, затрагиваемых в книге Агрести.

Упомянем также серию "Quantitative Applications in the Social Science", в рамках которой к настоящему моменту опубликовано более 120 брошюр (некоторые из которых упоминаются ниже). Работы рассматриваемого характера появляются и в рамках ряда других серий (названия известных нам серий даны в конце книги после списка использованной литературы).

Казалось бы, стоит перевести какие-то западные работы на русский язык - и проблема нехватки учебно-методической литературы в нашей стране будет решена. Однако, на наш взгляд, все не так просто. Конечно, перевод многих западных работ был бы весьма полезным для отечественной социологии (в частности, весьма полезным был бы перевод упомянутой выше работы Агрести). Но, как нам представляется, этого будет недостаточно.

Во-первых, западные описания отдельных методов (так же, как и отечественные) разбросаны по разным книжкам. И описания эти очень разношерстны в смысле степени пригодности для студентов-гуманитариев. Нам неизвестны книги, в которых был бы представлен некий минимальный набор методов, знание которых является необходимым каждому социологу. О соответствующих недостатках указанных выше учебников по математико-статистическим методам мы уже говорили. А, скажем, в той же книге Агрести, отсутствуют, к примеру, сведения об описательной статистике и т.д. (другими словами, в описываемых учебниках практически не рассматривается содержание второго раздела обсуждаемой второй части книги, а в книге Агрести не затронуто содержание первого раздела).

Во-вторых, предлагаемые социологу методы не сведены в систему. Авторы соответствующих работ не ставят в качестве цели формирование у исследователя-социолога

такого системного взгляда на характер решаемых задач и совокупность пригодных для этого методов, который мог бы послужить основой для формирования конструктивных алгоритмов выбора метода для решения той или иной конкретной задачи.

В-третьих, в западных работах, на наш взгляд, практически не уделяется внимания содержательному анализу моделей, заложенных в разных методах анализа данных, сравнительному изучению моделей, отвечающих алгоритмам, решающим сходные содержательные задачи.

В-четвертых, нам вряд ли стоит игнорировать отечественный опыт. Дело в том, что российскими учеными получено довольно много результатов, весьма полезных для социологии, лежащих в том же русле, что и некоторые западные алгоритмы, но имеющие определенные преимущества перед последними.

Во второй части книги мы в определенной мере пытаемся ликвидировать все указанные пробелы. В частности, в методическом плане излагаемое отличается следующими **особенностями**.

В *первом разделе* речь идет, в общем-то об известных вещах, много раз описанных в математико-статистической литературе. Но упор делается на те их аспекты, которые обычно остаются в стороне, несмотря на их важность для социолога: рассматриваются проблемы разбиения признаков на интервалы и работы с пропущенными данными, адекватность методов относительно типов шкал, специфика работы с дихотомическими данными, некоторые аспекты анализа моделей, предполагаемых используемыми методами.

Все методы, описанные во *втором разделе*, преподносятся как элементы единой системы, опирающейся на предлагаемую автором классификацию алгоритмов анализа связей. При описании каждого метода особое внимание уделяется анализу заложенной в нем модели. Модели, отвечающие разным методам, решающим одну и ту же задачу, сравниваются друг с другом. Обосновывается необходимость комплексного использования подобных методов.

Методические аспекты, затронутые во второй части книги, неотделимы от рассмотрений первой части.

Мы **предполагаем, что читатель знаком** с содержанием курсов по общей социологии, методике социологических исследований, теории измерений, математической статистике, предшествующих, в соответствии с принятыми в большинстве отечественных вузов (в том числе на социологическом факультете МГУ) учебными программами, курсу анализа данных. Рассмотрим коротко, что именно из указанных дисциплин должен знать читатель настоящей книги.

Что касается *курса общей социологии,* то на нем нам бы не хотелось останавливаться на нем подробно. Его освоение нужно просто для того, чтобы читатель понимал социальную значимость рассматриваемых в книге примеров. Другими словами, здесь речь идет об общей эрудиции читателя-социолога. Книг соответствующего профиля за последние годы вышло очень много. Мы их называть не будем, поскольку не это нас в первую очередь интересует.

Из *курса методики социологических исследований* прежде всего необходимо иметь представление об операционализации понятий, о видах исследований. О методике социологического исследования можно прочесть, например, в книге [Ядов, 1998].

Полагаем известными читателю подробно рассматриваемые в *курсе по теории измерений* определения основных типов используемых в социологии шкал: номинальной, порядковой, интервальной; сложности описания изучаемых объектов (в качестве которых чаще всего выступают респонденты) с помощью определенного набора признаков (отвечающим, например, вопросам в анкете), модельный характер такого описания; проблемы, связанные с получением от респондентов адекватной информации. С содержанием этого курса можно познакомиться по книге [Толстова, 1998а].

Считаем, что читатель имеет представление о роли *математической статистики* в социологическом исследовании: знает, что она изучает закономерности “в среднем”, дает возможность грамотно построить выборку и обобщить результаты с выборки на генеральную совокупность. Будем полагать также, что читателю известно хотя бы в самых общих чертах, что такое случайные величины, как они обычно бывают представлены в выборочной совокупности (когда вероятность какого-либо события отождествляется с относительной частотой его встречаемости, случайные величины отождествляются с признаками), знакомы основные принципы корреляционно-регрессионного анализа. Работ по теории вероятностей и математической статистике в отечественной литературе довольно много (как известно, отечественная наука в этом отношении имеет богатейшие традиции). Среди вышедших в последнее время и относительно "легких" в смысле преподнесения используемого математического аппарата можно назвать [Гмурман, 1998 а,б; Колемаев, Калинина, 1997]. Особенно хотелось бы отметить работу [Тюрин, Макаров, 1998], которая по своему достоинствам близка к названным выше западным учебникам и даже превосходит их более глубоким теоретическим обоснованием затрагиваемых методов.

Считаем также, что читатель знаком с *основными методическими принципами использования математики именно в социологии*: знает, что такое модель, заложенная в

математическом алгоритме; понимает суть органической связи между этапами измерения и анализа, важность решения проблемы однородности изучаемого массива данных; знаком со специфическими моментами интерпретации результатов анализа социологических данных. Об этом можно прочесть, например, в [Толстова, 1990а,б;1991а,б].

Конечно, когда в нашем изложении встретится необходимость использования какого-либо из названных положений, мы будем коротко напоминать его читателю. О многих положениях речь пойдет довольно подробно (особенно это касается первой части). Но, тем не менее, априорное знание этих положений читателем очень желательно, поскольку мы не ставим своей целью излагать их так, как этого требует жанр учебного пособия. Скорее мы претендуем на сведение названных положений в некоторую "социолого-математическую" систему. Это касается обеих частей книги. Первой – поскольку она полностью посвящена методологии статистического анализа социологических данных. Второй – в силу того, что мы не просто описываем наиболее актуальные для социолога методы, а предлагаем их определенную систематизацию, опирающуюся не некоторое методологическое видение задач эмпирической социологии.

## *Часть 1.*

## *ЧТО ТАКОЕ АНАЛИЗ ДАННЫХ?*

## *(Методологический аспект)*

### 1. ПОИСК СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ КАК ОСНОВНАЯ ЦЕЛЬ, СТОЯЩАЯ ПЕРЕД ЭМПИРИЧЕСКОЙ СОЦИОЛОГИЕЙ. РОЛЬ АНАЛИЗА ДАННЫХ В ЕЕ ДОСТИЖЕНИИ

#### 1.1. Эмпирическая основа для изучения социальных явлений

Роль эмпирических данных в изучении социальных явлений огромна. Достаточно глубокое изучение интересующих социолога закономерностей невозможно без опоры на анализ конкретных фактов, в которых эти закономерности, собственно говоря, и проявляются. "Питательной" средой для теоретических построений чаще всего является эмпирический материал[1]. Именно реальные эмпирические факты[2], как правило, служат средством проверки теорий, наводят на мысль о необходимости их корректировки, служат почвой для формирования новых теоретических гипотез.

Что же такое социологические эмпирические данные, т.е. данные, характеризующие конкретные социологические факты; данные, в виде которых, собственно говоря, эти факты перед нами и выступают? Данные могут представать перед исследователем в виде:

- совокупности чисел[3,] характеризующих те или иные объекты (в качестве таких совокупностей могут выступать, например, производственные характеристики предприятий, возраст респондентов, оценки выпускниками школ престижности некоторых профессий и т.д.)[4,]

- множества индикаторов определенных отношений между рассматриваемыми объектами (к примеру, при изучении производственных бригад такими индикаторами могут служить указания каждого члена бригады на то, нравится ли ему работать вместе с любым другим членом той же бригады, такие данные часто используются при изучении малых групп [Математические методы анализа ... 1989, гл. 4]),

- результатов попарных сравнений респондентами каких-либо объектов (такие данные используются в методе парных сравнений [Дэвид, 1978] - способе построения шкал, отражающих усредненное отношение изучаемой совокупности респондентов к каким-либо объектам).

- совокупности определенных высказываний (например, ответов респондентов на вопрос об их профессии, о том, что им нравится в политике правительства; письма читателей газеты в редакцию; фрагменты из журнальных статей и т.д.),

- текстов документов;

- так или иначе зафиксированных результатов наблюдения за невербальным поведением каких-либо людей и т.п.

Наиболее часто в социологических исследованиях данные представляют собой совокупность значений каких-либо признаков (характеристик, переменных, величин; будем считать эти термины синонимами), измеренных для каждого из изучаемых объектов.

Мы не будем глубоко анализировать смысл термина "признак", хотя здесь есть о чем поговорить (на наш взгляд, это понятие требует специального обсуждения; здесь мы такой цели перед собой не ставили). Будем считать этот смысл в основном интуитивно ясным. Отметим лишь некоторые моменты.

Признак - это некоторое общее для всех объектов качество, конкретные проявления которого (значения признака; их называют также альтернативами, градациями), вообще говоря, могут меняться от объекта к объекту. Примеры признаков - пол, возраст респондентов, их удовлетворенность своим трудом и т.д. В качестве значений признака "возраст" могут выступать 25 лет, 48 лет, 21 год. Для нас важно, что само введение практически любого признака является моделированием довольно высокого уровня. Признаки не существуют сами по себе, они - плод наших абстрактных рассмотрений, идеальные конструкции. В общественных науках соответствующий процесс абстрагирования является иногда очень непростым. Основными его этапами является выделение понятий (процесс рождения которых уже не прост[5]) и осуществление их т.н. операционализации. Процессу операционализации понятий посвящена обширная литература[6]. Мы не будем описывать то, что читатель может из нее почерпнуть. Отметим лишь, что, на наш взгляд, его надо понимать несколько шире, чем это обычно делается. Так, в него имеет смысл включить, например, различные способы шкалирования (скажем, получение на основе непосредственного опроса респондентов значений некоторых вспомогательных признаков и последующий переход к другим, латентным переменным с помощью построения индексов, как это делается, например, при построении известной шкалы Лайкерта).

На практике проблему операционализации чаще всего разделяют на две: выбор признаков, являющихся индикаторами понятий, и выбор набора значений каждого признака (скажем, выбрав в качестве одного из индикаторов признак "возраст", мы можем считать его

"непрерывным" и просить каждого респондента указывать целое число прожитых лет; а можем – приписывать респонденту число от 1 до 5 в зависимости от того, в какой возрастной интервал респондент попадает: от15 до 25 лет, от 25 до 35 лет, …, старше 55 лет; вполне возможно, что мы разделим всех людей лишь на две группы – до 30 лет и старше и т.д.). Ниже (п.1.3) покажем, что в процесс операционализации имеет смысл включить также процедуру определения типа используемых при получении значений наблюдаемых признаков шкал. Покажем также, что этот процесс не может осуществляться в отрыве от анализа данных и интерпретации его результатов.

При концептуализации понятий должны решаться вопросы, отнюдь не лежащие на поверхности. Напротив, успешная операционализация предусматривает переход на достаточно глубокий концептуальный уровень рассмотрения предмета исследования, при котором признаки воспринимаются как отражение параметров анализа, релевантных целям исследования, а значения признаков - как результат расчленения каждого параметра на определенные категории, ключевые понятия исследования.

Подчеркнем также, что, как известно, при получении информации от респондента огромную роль играет не только сам перечень градаций-ответов на вопросы анкеты, но и порядок упоминания этих градаций, конкретный выбор слов при их формулировке, преамбула к вопросу, порядок вопросов в анкете и т.д. (см., например, [Мосичев, 1996; Questions and answers …, 1996]). Обо всем этом мы говорить не будем, неявно имея в виду необходимость решения соответствующих проблем.

Вопрос о самом существовании признака, о трактовке его значений бывает иногда очень тонким (см., например, работу [Ноэль Э., 1993], автор которой, несмотря на сугубо практическую направленность книги, считает нужным оговорить соответствующие теоретические вопросы, вводит понятие "мышление признаками" и анализирует плюсы и минусы перехода к такому мышлению).

Далее будем рассматривать ситуацию, когда каждый изучаемый объект предстает перед нами в виде последовательности чисел – значений для него неких признаков. Такие данные обычно задаются в виде таблицы (матрицы) "объект-признак", строки которых отвечают объектам (например, респондентам), а столбцы – признакам (например, каждый столбец – это ответы респондентов на один из вопросов анкеты). Пример такой таблицы представлен ниже.

*Таблица 1*

**Пример таблицы "объект-признак"**

| Номер объекта (респондента) | Наименование признака | | |
|---|---|---|---|
| | Пол (0 – муж., 1 - жен.) | Возраст (лет) | Удовлетворенность трудом (1- совершенно не удовлетворен,…, 5- полностью удовлетворен) |
| 1 | 0 | 25 | 1 |
| 2 | 0 | 31 | 2 |
| 3 | 0 | 18 | 5 |
| 4 | 1 | 24 | 2 |
| 5 | 0 | 18 | 1 |
| 6 | 0 | 38 | 4 |
| 7 | 1 | 41 | 3 |
| 8 | 1 | 50 | 1 |
| 9 | 1 | 54 | 2 |
| 10 | 1 | 19 | 5 |

При использовании методов многомерного анализа данных ту же информацию об исходных объектах зачастую представляют в виде фрагмента так называемого признакового пространства: осям такого пространства отвечают рассматриваемые признаки, а каждый объект представлен в виде точки, координатами которой служат значения для этого объекта признаков, отвечающих осям. Ниже приведен пример двумерного признакового пространства (рис.1),

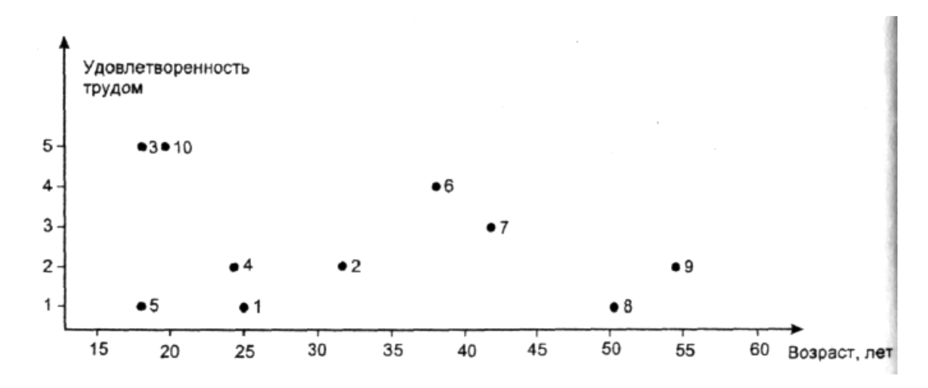



**Рис. 1. Пример двумерного признакового пространства.**

Отмеченные точки отвечают респондентам, координаты которых заданы таблицей 1

оси которого отвечают признакам "возраст" и "удовлетворенность трудом", а координаты объектов отвечают данным таблицы 1.

Подчеркнем, что подобное представление изучаемых объектов, будучи исходным для алгоритмов анализа данных, в действительности скрывает (должно скрывать!) за собой глубокую предварительную работу исследователя по осмыслению того, что и почему он изучает (несколько более подробно мы рассмотрим это положение в п. 1.3). На этот принципиальный момент обращают внимание многие авторы. Например, Чесноков говорит о глубокой принципиальной значимости матрицы "объект-признак". Батыгин пишет о том. что "…трехкомпонентная логико-семантическая структура, включающая объект, переменную и ее значение, составляет своеобразный … формат организованного знания, образующий привычную для социолога матрицу данных" [Батыгин, 1986, с. 135].

Итак, перед нами стоит некоторая социологическая задача и мы полагаем, что для ее решения необходимо изучить определенное количество данных о некоторых объектах. Например, предположим, что перед нами лежит 1000 заполненных анкет, в каждой из которых фигурирует 50 обращенных к респонденту вопросов[7] . Допустим, что мы догадываемся о том, что в этих данных скрываются интересующие нас закономерности (полагаем, что вопросы, включенные в анкету, были тщательно продуманы, увязаны со сформулированными заранее гипотезами исследования и т.д.). Но как их "выудить" из того огромного количества цифр, которые имеются в нашем распоряжении? Как не "потеряться" в этом море информации? Как "продраться" сквозь все эти необозримые данные, суметь увидеть то, что нас интересует? Заметим, что проблема поиска способа "плавания" по описанному "морю" встает, отнюдь, не только перед таким исследователем, который не знаком с методами анализа данных. Дело в том, что специфика, сложность социальных явлений приводит к многочисленным трудностям анализа, вызывает необходимость весьма творческого подхода к его осуществлению. Об этом и пойдет речь ниже.

## 1.2. Понятие статистической закономерности. Роль статистических и нестатистических закономерностей в эмпирической социологии

В науке принято выделять две основные формы закономерной связи явлений, отличающиеся по характеру вытекающих из них предсказаний: динамические и статистические закономерности (см., например, [Философский энциклопедический словарь,1983. С.653]). В законах динамического типа предсказание имеет точный, определенный однозначный вид; в

статистических же законах предсказание носит не достоверный, а лишь вероятностный характер[8].

Ниже нас будут интересовать в основном статистические закономерности (поиск таких закономерностей - основная цель анализа данных). Это - закономерности "в среднем".

Авторы книги [Тюрин, Макаров,1998, с. 18] пишут о том, что статистический подход состоит в мысленном разделении наблюдаемой изменчивости на две части, обусловленные, соответственно, закономерными и случайными причинами, и выявлении закономерной изменчивости на фоне случайной. Вероятностный характер предсказаний в статистических закономерностях обычно бывает обусловлен действием множества случайных факторов, которые имеют место в статистических совокупностях. Статистическая закономерность возникает как результат взаимодействия большого числа элементов, составляющих совокупность и характеризуют не столько поведение отдельного элемента совокупности, сколько всю совокупность в целом. Проявляющаяся в статистических закономерностях "необходимость" возникает вследствие взаимной компенсации и уравновешивания множества случайных факторов, "пробивает" себе дорогу через массу случайностей, контрпримеров, отступлений от нее.

Другими словами, интересующий нас подход позволяет "за деревьями увидеть лес" – например, за специфичностью, неповторимостью каждого человека усмотреть тенденции, имеющие место "в среднем" для всех респондентов изучаемой совокупности.

Статистическими являются часто употребляемые социологами утверждения типа: "средний возраст рабочих-металлургов равен 30 годам", "выбор профессии выпускниками школ не связан с их полом", "такая-то радиопередача имеет самый высокий рейтинг среди слушателей" и т.д.

Роль изучения статистических закономерностей для социологии вряд ли можно переоценить. Они вполне адекватно описывают массовые явления случайного характера, а именно такого рода явления и изучает обычно социолог.

О громадной роли изучения статистических закономерностей в эмпирических науках, в том числе – в эмпирической социологии, говорят многие авторы – философы, социологи, историки, математики. См., например, [Ракитов, 1981; Давыдов, 1991; Ноэль, 1993; Гумилев, 1993; Тюрин, Макаров, 1995; Фелингер, 1985][9].

Несмотря на сказанное, в литературе нет единого понимания и смысла, и роли статистических закономерностей в социологии. Поскольку понятие статистической закономерности является для нас ключевым, более подробно рассмотрим некоторые

представляющиеся нам принципиальными аспекты, связанные с пониманием статистического подхода именно социологами.

Чаще всего, говоря о статистичности социальных закономерностей, исследователи имеют в виду законы развития больших социальных групп и общества в целом. При этом подобная статистичность обычно рассматривается в контексте анализа известной дилеммы о соотнесении общих закономерностей развития общества и свободой воли отдельного человека.

Например Л.Н. Гумилев, излагая свою известную теорию этногенеза, описывая конкретные исторические процессы, неоднократно подчеркивает, что "зигзаги истории" погашаются "статистической закономерностью этногенеза" [Гумилев, 1993, с.634], "там, где царит вероятность, детерминизм неуместен" [там же, с. 654], "статистический ход событий выше возможностей одного человека" [там же, с. 660] и т.д.

В работе [Никитина, 1996] приводится следующая цитата из Э.Маха (модного ныне; его творчеству в нашей литературе сейчас уделяется довольно интенсивное внимание в связи с интересом к истории позитивизма): "В статистике действительно применяется метод исследования, основанный на намеренном пренебрежении, игнорировании индивидуального в изучении наиболее существенных, наиболее между собою связанных обстоятельств. И действительно, при этом произвольные действия людей оказываются в такой же мере закономерными, как какой-нибудь растительный и даже механический процесс, при котором никто обыкновенно не думает о психическом воздействии, о влиянии воли. Число браков и самоубийств в течение года в какой-нибудь стране колеблется столь же мало, если не еще меньше, как число рождений и случаев естественной смерти, хотя в первых воля играет как будто большую роль, а в последних – никакой." Как известно, аналогичные высказывания имеются и у известных социологов – классиков – например, у Кондорсе, Дюркгейма, Парка, Парето (отметим, что статистическое понимание исследуемых закономерностей социологами-позитивистами, начиная с Конта, сознательно связывается с отказом от изучения причинно-следственных отношений: вместо причинности – статистические связи).

Симптоматичным представляется то, что в последние годы в работах специалистов по теоретической социологии стали появляться параграфы с названиями типа: "Программа статистически-вероятностно ориентированной науки об обществе" (о творчестве Кондорсе) [Давыдов, 1995].

Нахождение разного рода статистических закономерностей является привычным делом каждого социолога, проводящего эмпирическое исследование. Но нам представляется некорректным, когда статистический подход связывается только с большими группами или

обществом в целом, что явно делается и в приведенных выше цитатах, и (явно или неявно) в работах многих других авторов. И за подобными суждениями стоит нечто весьма принципиальное.

Поясним нашу точку зрения, обратившись к представляющейся очевидной связи процитированных выше рассуждений с известным обсуждением соотношения принципов номинализма и реализма в социологии. Если учесть, что номинализм иногда (и не совсем корректно) ассоциируется с т.н. гуманитарной парадигмой, то отсюда – один шаг до противопоставления статистического и гуманитарного подходов. Так и поступают многие авторы. Для нас такой взгляд совершенно неприемлем. По нашему мнению, указанные термины отвечают различным основаниям выделения исследовательских подходов. Статистические модели могут использоваться и при попытке "понять" отдельного человека (а гуманитарный подход, как известно, близок к подходу "понимающей" социологии), и при изучении разного рода групп людей, в том числе общества в целом. Пример применения статистического подхода к изучению представлений отдельного человека – использование статистических распределений для описания неоднозначности мнения одного респондента относительно любого объекта. Такой подход используется во многих методах шкалирования (например, в Терстоуновской модели метода парных сравнений [Клигер и др., 1978; Суппес и Зинес, 1967; Толстова, 1998а], в известной модели Рашевского подражательного поведения [Математические методы…, 1966; Моделирование социальных …, 1993]). И его появление явилось следствием именно попытки "понимания" того, что происходит в сознании отдельного индивида (вероятно, здесь будет кстати упомянуть, что без хотя бы какой-то формализации изучаемого явления никакое его научное изучение, невозможно; об этом мы коротко говорили в [Толстова, 1996б]).

С нашей точки зрения, адекватно отражающими суть статистического подхода при изучении отдельного человека являются соображения, приведенные в работе [Фелингер, 1985], где актуальность для социологии изучения статистических закономерностей аргументируется посредством рассмотрения детерминированной и стохастической (вероятностной) составляющей в психологии человека, анализа механизма выполнения эмоциональными формами психологической деятельности человека роли стохастических регуляторов поведения. Примерно о том же говорится в [Давыдов, 1991].

Вероятно, непросто добиться того, чтобы представление о статистичности мнения одного респондента многих закономерностей для больших групп респондентов далеко не сразу была воспринята человечеством. Интересный исторический экскурс осуществляет, например Э. Ноэль в своей книге [Ноэль Э., 1993]. Она описывает процесс становления статистического

подхода к изучению социальных процессов, показывает, что негативное отношение к такому подходу имеет долгую историю (уже "в Ветхом завете есть указание на то, что применение статистики к людям следует считать опасным" [там же, с. 20]; но и в наше время к собиранию данных с помощью выборочного метода многие относятся "как к трюку фокусника" [там же, с. 18] (выборочный метод как известно, – сердцевина статистического подхода, см. разделы 3, 4).

Анализ данных позволяет находить статистические закономерности. Этим определяется его важность для социолога. Но чтобы в полной мере оценить его роль в эмпирической социологии, попытаемся ответить на противоположный вопрос: в какой мере социолог может ограничиться поиском закономерностей "в среднем"? Вряд ли значимость анализа данных для социологии можно оценить в полной мере, не зная, от чего мы отказываемся, ограничиваясь использованием только статистической парадигмы[10].

Отвечая на поставленный вопрос, выделим два аспекта. В обоих случаях, как мы увидим, речь по существу пойдет о рассмотрении изучаемых явлений в рамках системной парадигмы (мы без объяснения используем терминологию системного анализа, надеясь на интуитивное понимание читателем соответствующих терминов)[11].

Во-первых, весьма важным для социологии является поиск динамических закономерностей. Мы не будем строго их определять. Ограничимся лишь упоминанием того, что в результате поиска таких закономерностей строятся модели мобильности в социальных системах, модели процессов межличностного влияния и внутриличностных конфликтов, модели подражательного поведения и т.д. [Бартоломью, 1985; Моделирование социальных ..., 1993; Паповян, 1983; Плотинский, 1992, 1998]. Соответствующие методы обычно называют методами моделирования социальных процессов (название не совсем удачное с точки зрения противопоставления этих методов методам анализа данных, поскольку последние - это тоже методы построения определенного рода моделей; по существу все изложенное ниже можно расценивать как обсуждение проблем, связанных с обеспечением их адекватности). Методы моделирования часто опираются на расчет дифференциальных уравнений, отражающих скорость изменения того или иного процесса, либо на матричную алгебру[12].

Во-вторых, социолога не могут не интересовать и многие такие явления, которые не носят статистического характера по несколько иным причинам. Поясним это, отталкиваясь от рассмотренных выше примеров статистических закономерностей. Скажем, может быть интересно выяснить, каким образом в качестве рабочих-металлургов, средний возраст которых равен 30 годам, "умудряются” функционировать отдельные люди старше 60 лет; почему при отсутствии статистической связи между полом выпускника школы и выбором им профессии на

социологический факультет МГУ в этом году поступили практически одни девушки; в чем причина того, что, несмотря на общий высокий рейтинг передачи, одна из зрительниц прислала на радио письмо с резко отрицательной ее оценкой и т.д.

Интересуясь лишь статистическими закономерностями, мы игнорируем "аномалии", отступления от средней зависимости, что вряд ли можно считать допустимым. Заметим, что анализ "аномалий" предусматривается грамотным использованием традиционных статистических методов: любой статистический пакет предусматривает выдачу пользователю т.н. резко отклоняющихся наблюдений. Но такие специфичные объекты могут интересовать социолога не только как некоторые "огрехи" найденной статистической закономерности. Вряд ли следует бездумно выбрасывать соответствующие объекты из дальнейшего рассмотрения. Иногда (скажем, в критические моменты развития общества) анализ мнения такого "отклоняющегося" респондента может дать больше, чем выявление многих статистических закономерностей (может быть, имеющих место только в силу инерции, обреченных на исчезновение в ближайшем будущем).

Изучение фактов, не укладывающихся ни в какие статистические закономерности, анализ случайных флуктуаций, не выражающихся в статистически значимых характеристиках, может стать целью исследования. Исследователь может стремиться найти такие "возмущения" в общественной жизни, такие ее "переломные" точки системы, которые свидетельствуют либо о ее разрушении, либо о зарождении новой системы.

Естественно, что при такой постановке задачи методы, направленные на поиск "средних" закономерностей, скрывающихся за наблюдаемыми фактами, т.е. статистические методы, перестают играть главенствующую роль.

Поиск уникальных точек вообще может не ассоциироваться с поиском закономерности. Для пояснения этого положения вспомним восходящее к Виндельбанду и Риккерту разделение всех наук на номотетические и идиографические (олицетворением которых считают обычно, соответственно, физику и историю; подчеркнем, что слово "идиографический" происходит от греческого ιδιοσ - своеобразный, странный необычный, а не от ιδεα - форма постижения в мысли явлений объективной реальности; поэтому писать это слово следует через букву "и" [Давыдов, 1986]). Первые рассматривают действительность с точки зрения всеобщего, выражаемого с помощью некоторых законов. Вторые – образные науки, описывающие единичное в его эмпирической неповторимости. Вероятно, "в чистом виде" ни те, ни другие науки не встречаются. Самая "номотетическая" наука ищет закономерности, опираясь в

конечном счете на изучение уникальных объектов. И, напротив, любая "наиидиографичнейшая" наука все-таки пытается в той или иной мере "выйти" на общие закономерности (с нашей точки зрения, наука начинается там, где в разных объектах исследователь начинает находить что-то общее, т.е. там, где уникальность, неповторимость объектов исчезает). Другими словами, понятия идиографической и номотетической науки сродни веберовским идеальным типам (к слову заметим, что, вероятно, то же можно сказать и относительно так называемых социологических номинализма и реализма).

Социология же в принципе находится "между двух стульев". Все выявляемые закономерности слишком приблизительно отражают то, что интересует исследователя. Поэтому для социолога очень остро стоит вопрос о постоянном неформальном изучении отдельных объектов (в первую очередь, - людей), о своего рода мониторинге в деле неформального отслеживания специфики изучаемых явлений. В частности, актуальным является поиск объектов, не похожих на других, уникальных в своем роде. И существуют методы, позволяющие это делать. В число наших задач не входит их описание. Однако коротко упомянем некоторые из них, более ясно "очертив" тем самым границы анализа данных.

Вероятно, для социолога наболее важными методами, позволяющими находить и изучать уникальные точки рассматриваемой системы, являются т.н. мягкие методы общения с респондентами и анализа полученной от них информации. В литературе нет установившейся традиции по поводу четкой трактовки этого термина [Ядов, 1991]. Скажем лишь, что к числу мягких методов опроса относятся, например, биографический метод, разные виды не формализованного интервью – глубинное, фокусированное (в том числе - групповое, или метод фокус-групп), с путеводителем, лейтмотивное, полуформализованное и т.д. К мягким методам анализа можно отнести некоторые методы работы с текстами. Можно говорить о мягкости всей стратегии исследования (см. упомянутую выше работу Ядова). В таких случаях говорят о качественной социологии. (Правда, необходимо подчеркнуть, что в литературе ведется очень много споров по поводу понимания этого термина и целесообразности его введения в науку [Батыгин, Девятко, 1994].) Подробнее о мягких методах можно прочитать, например, в работах [Семенова, 1998; Ковалев, Штейнберг, 1999]. Это – первые отечественные учебники по качественной социологии, хотя на западе таких учебников довольно много. В западной социологии в последние десятилетия бурно развиваются способы анализа данных, полученных с помощью мягких методов (соответствующие данные называют качественными). Родился новый термин – “анализ качественных данных”, отражающий мощное направление изучения текстов (в виде которых обычно предстают перед исследователем результаты мягких способов общения с

респондентом). Его роль и статус близки к роли и статусу интересующего нас "анализа данных". Поскольку в России "анализ качественных данных" практически неизвестен, мы позволили себе привести в библиографии список "Учебники по анализу качественных данных". Это направление эмпирической социологии весьма перспективно и уже настолько четко оформилось, что в нем очень активно используется математика (см. в библиографии список "Математические методы в качественной социологии"). [13]

Ясно, что мягкие методы действительно дают возможность изучать уникальные настроения отдельных респондентов (см., например, [Ярская-Смирнова, 1997]) и тем самым подводить исследователя к обнаружению "точек перелома" системы. Заметим, однако, что они же могут использоваться и для прямо противоположных целей - для поиска закономерностей "в среднем". Скажем, для изучения мнений самых типичных, "средних" респондентов с целью "ориентации" социолога в новой для него проблеме, для более успешной формулировки гипотез, вопросов в анкете (например, в процессе пилотажного исследования) и т.д.

Мягкие методы позволяют "докопаться" до истины не на основе каких-либо формальных схем, а с помощью творческого использования интеллекта, опыта, интуиции исследователя. Однако и здесь может активно применяться математика, в том числе и статистические методы, хотя при этом речь идет не о "среднем" для группы людей, а о "среднем" мнении отдельного респондента; поиск такого "среднего" тоже имеет смысл, поскольку и у отдельного человека "истинное" мнение может искажаться случайными факторами.

Несколько слов - о математических методах, направленных на поиск уникальных объектов. Наиболее известные методы такого рода относят обычно к упомянутым методам моделирования социальных процессов. Примером могут служить методы синергетики, представляющиеся весьма актуальными для социологии, но крайне редко использующиеся российскими исследователями [Бранский, 1997; Евин, Петров, 1991; Капица и др., 1997; Князева, Курдюмов, 1994; Курдюмов и др., 1989]. Ниже будем считать, что нас интересуют только статистические закономерности (хотя многие приводимые ниже положения имеют отношение не только к ним).

## 1.3. Проблема соотнесения формального и содержательного при формировании представлений о закономерности в социологии

Ясно, что используя математический аппарат для решения тех или иных практических

задач, мы всегда имеем дело не с самой реальностью, а лишь с некоторой ее моделью. А модель уже в силу того, что она модель, не может не содержать элементов формализации реальности. В какой-то степени это очевидно. И если бы речь шла об естествознании или технике, то мы не стали бы здесь говорить о заявленной в заголовке проблеме. Для удовлетворения потребностей естественных наук были разработаны такие методы, которые опираются на модели, в достаточной степени похожие на реальность (или, во всяком случае, достаточно хорошо отражающие представления исследователя об этой реальности). Это подтверждается тем, что прогнозы, осуществляемые на основе использования математических методов, обычно оправдываются. Другими словами, степень приближения модели к реальности оказывается достаточной для удовлетворения потребностей практики. Так, строитель при постройке дома рассчитывает нагрузку на какую-то балку, не задумываясь о том, что "работает" при этом не с самой балкой, а с некоторой ее формульной моделью. Более или менее ясно, что с чем здесь соотносится, и грамотно выполненные расчеты обеспечивают эффективность соответствующего модельного подхода.

Не та ситуация в социологии (и других общественных науках). Сложность соответствующих явлений влечет сложность формализации наших представлений о них. Модели реальности, которые мы фактически строим, используя тот или иной метод анализа данных, оказываются чересчур приблизительными, соответствующие прогнозы не сбываются и т.д. Эти модели настолько субъективны, что исследователь все время рискует получить результаты, плохо отражающие реальность. Поэтому он должен постоянно отслеживать, какой моделью вольно или невольно пользуется, думать о соотнесении формального и содержательного.

И начинается этот процесс с формирования самых первичных, зачастую весьма смутных, представлений социолога о том, что он, собственно, должен изучать. Ниже мы попытаемся описать подобные начальные шаги. При этом ограничимся рассмотрением лишь некоторых принципиальных моментов – таких, без учета которых немыслим эффективный анализ социологических данных и, прежде всего, - выбор метода анализа. Подчеркнем, однако, для действительно успешного анализа необходимо более глубокое изучение вопроса; здесь очень много не решенных проблем.

Заметим, что, в соответствии с логикой построения настоящей работы, все сказанное ниже в настоящем параграфе нужно было бы отнести в раздел 5, специально посвященный специфике поиска статистических закономерностей именно в социологии. Там и пойдет речь о чем-то схожем – о "приспособлении" известного формализма анализа данных к конкретной

исследовательской ситуации; а здесь – в основном о том, как в сознании социолога рождается сама потребность прибегнуть к формализму.

Итак, мы предполагаем, что общество развивается в соответствии с некоторыми закономерностями, на изучение которых и направлены интересующие нас действия социолога. Судить же об этих закономерностях он может только на основе имеющихся в его распоряжении данных, которые можно расценивать как результаты измерения (заметим, что, когда такими результатами служат числа, вместо термина "измерение" часто используют термин "шкалирование"), как модель (чаще всего – математическую) реальности. Начнем с рассмотрения основных принципов построения этой модели. По существу речь пойдет о некоторых аспектах формирования и операционализации понятий. Можно также сказать, что мы коснемся *проблемы интерпретации данных*, подлежащих анализу.

Прежде всего, "усмотрим" в исходных данных как бы два уровня: множество скрывающихся за ними реальных объектов (отдельных людей, социальных групп, институтов и т.д.) во всей их уникальности и неповторимости и получающуюся в результате непосредственного сбора данных совокупность отражающих эти объекты формальных конструктов: чисел, текстов и т.п. Описанные уровни можно расценивать как *содержательный и формальный аспекты данных*. Сразу подчеркнем, что термин "содержательный" здесь употреблен в значительной мере условно: когда исследователь приходит к выводу о необходимости изучать именно такие-то объекты, он уже имеет в своем сознании некоторые, иногда весьма сложные и всегда – субъективные, представления о том, почему он это делает; и эти представления бывают основаны на том, что в объектах усматривается нечто общее, т.е. на отказе от их "уникальности и неповторимости" (обычно это общее выражается в описании всех объектов значениями каких-то выбранных исследователем признаков). Детальное изучение истоков такого происходящего в сознании человека процесса абстрагирования от реальности не входит в наши задачи. Отметим лишь, что этот процесс не отделим от формирования представлений об объекте и предмете исследования (надеемся, что читателю очевидно различие трактовок термина "объект" в сочетаниях "реальные изучаемые объекты" и "объект исследования"; ср. сноску [4]). Будем полагать, что совокупность соответствующих априорных представлений социолога, не предполагающих не только абстрагирования от уникальности изучаемых объектов, но и, может быть, самого вычленения этих объектов (предполагается, однако, что в дальнейшем эти представления будут служить базой для такого вычленения), образуют фрагмент *априорной содержательной модели* (второй фрагмент этой модели связан с априорными представлениями исследователя об изучаемых закономерностях, он рассмотрен

ниже).

Содержательный и формальный уровни исходных данных отвечают определенным этапам процесса измерения. При анализе данных мы используем последние в их формальном виде. Но эффективный анализ может быть осуществлен лишь на основе грамотного соотнесения формального аспекта данных с содержательным, более того, - с соответствующим фрагментом априорной содержательной модели. Задумавшись же о том, каким образом можно перейти от содержательных рассмотрений к формальным, мы наверняка придем к выводу, что существует еще один, промежуточный, этап процесса измерения. Он отвечает тому логическому вычленению в многоцветной реальности, ассоциируемой с предметом исследования, и изучаемых объектов, и их отдельных сторон, которое связано с формированием и операционализацией понятий, т.е. с выбором конкретных объектов измерения и способов сбора данных. Этот этап можно считать фрагментом построения *концептуальной модели* реальности (второй фрагмент этой модели связан с выбором алгоритма анализа результатов измерения и будет рассмотрен ниже)[14].

Отметим, что в действительности вопрос о построении концептуальной модели, отвечающей процессу измерения очень сложен. Формируя понятия, лежащие в основе наших представлений о виде измеряемых признаков, мы должны решать множество вопросов о взаимопонимании респондента и исследователя, о том, каким образом опрашивать людей (скажем, выбрать степень “жесткости” опроса), задействовать или нет те или иные “хитрые” способы шкалирования и т.д. При использовании “жестких” методов необходимо определить точный набор значений измеряемых признаков, расположение соответствующих вариантов ответов в анкете, структуру преамбулы к вопросу и т.д. Применяя “мягкие” методы, необходимо решить огромное количество весьма сложных вопросов, связанных с кодированием получаемых текстов, усматриванием общих свойств у разных респондентов (чтобы перейти к анализу данных, необходимо перейти к “мышлению признаками”).

Здесь же – определение тех объектов, для которых будет непосредственно осуществляться измерение (построение и корректировка выборки), решение ряда проблем, связанных с реализацией процедуры измерения (например, учет влияния интервьюера на результат опроса) т.д. и т.п.

Отметим также, что именно на этапе построения концептуальной модели рассматриваются вопросы, связанные с построением эмпирической и математической систем, о которых пойдет речь в п. 2.2. Соответствующие рассмотрения – это взгляд на весь процесс концептуализации с другой, пока не использованной нами точки зрения – точки зрения теории

измерений. Этот взгляд является необходимым, если исследователь хочет обеспечить адекватность используемого для анализа данных математического аппарата характеру решаемой социологической задачи.

Реализация выбранных способов сбора данных приводит нас к фрагменту *формальной модели* реальности (второй фрагмент будет получен в результате реализации метода анализа данных).

Итак, в процессе интерпретации подлежащих непосредственному анализу формальных данных мы выделили их содержательный, концептуальный и формальный (как правило, - математический) аспекты. Они отвечают построению априорной содержательной, концептуальной и формальной модели реальности в процессе измерения. Аналогичные аспекты можно выделить и в понимании искомой закономерности. Попытаемся это сделать.

Как отмечалось в п.1.1, именно в качестве исходных данных (здесь добавим - в их формальном виде) выступают перед исследователем те факты, характер которых объясняется действием искомых закономерностей. Другими словами, эти закономерности как бы являются "причинами" того, что наши факты имеют заданный вид. Скажем, если наши формальные данные, набор фактов – это измеренные для ряда регионов страны уровни безработицы и число суицидов на 1000 жителей, то специфический характер этих фактов может состоять, например, в том, что с ростом безработицы, как правило, наблюдается увеличение доли суицидов; а "причина" такого вида фактов – в том – что материальная необеспеченность людей толкает их к самоубийству[15]. Однако эти "причины" остаются для нас латентными. В явном виде они выступают перед нами как закономерности другого рода – некие формальные соотношения, связывающие отдельные элементы формальных же данных друг с другом. В нашем примере это может быть близость к единице коэффициента корреляции между используемыми переменными.

Первые закономерности назовем *содержательными* (термин "содержательный" здесь тоже можно использовать лишь условно, и степень условности – еще б**о**льшая, чем условность использования того же термина в выражении "содержательный аспект исходных данных"; мы огрубляем ситуацию; уже само использование исследователем понятия "закономерность" означает наличие в его сознании некой модели), вторые – *формальными*. Можно сказать, что формальная закономерность служит для нас статистическим подтверждением правильности нашего предположения о существовании содержательной закономерности. Представления о содержательных закономерностях являются вторым фрагментом упомянутой выше априорной содержательной модели. Найденные же в результате анализа данных формальные

закономерности – вторым фрагментом формальной модели.

Нетрудно видеть, что между содержательной и формальной закономерностью тоже стоит некоторая концептуальная модель реальности. Во всем многоцветье реальных взаимодействий наблюдаемых объектов друг с другом мы вычленяем соотношения, которые называем, к примеру, наличием связи между рассматриваемыми понятиями (имея в виду понятия, выделенные при построении концептуальной модели, задействованной в процессе измерения). Эти соотношения должны, в частности, дать нам основания для выбора конкретного способа анализа данных, конкретного формализма, отвечающего постановке нашей содержательной задачи. Подобные соотношения имеет смысл назвать *концептуальной моделью* изучаемой закономерности. Таким образом, выбор метода – часть построения концептуальной модели искомой закономерности.

Процесс концептуализации представлений социолога об искомых закономерностях нельзя оторвать от построения описанной выше “измерительной” концептуальной модели. Само понимание закономерности непосредственным образом замыкается на то, какие понятия мы выбрали для изучения, как их операционализировали и т.д. Ниже будем говорить о построении единой концептуальной модели реальности, предшествующем анализу данных (точнее, являющемся его не всегда осознаваемой частью). В аналогичном, "объединительном", смысле будем использовать термины "априорная содержательная модель" и "формальная модель".

Чтобы логически завершить наши рассуждения, отметим, что выбором и реализацией конкретного алгоритма анализа данных работа социолога по поиску интересующих его закономерностей, конечно, не кончается (заметим, что мы здесь не говорим об этапе непосредственной реализации метода, поскольку здесь социолог не выступает именно как социолог). Далее наступает этап интерпретации результатов применения алгоритма. Зачастую этот этап бывает сложным, требующим весьма неординарного искусства социолога. Только в результате реализации интерпретационного этапа мы получим представление о “причинах”, упомянутых выше. И, вероятно, достаточно корректный анализ действительных причин не может осуществляться, помимо всего прочего, без использования качественных методов. Получив высокое значение коэффициента корреляции между уровнем безработицы и количеством суицидов в регионе, мы вряд ли будем уверены в объективности соответствующих выводов причинно-следственного характера, если не прибегнем к серьезному изучению поведения отдельных людей, страдающих от безработицы. Только качественные методы могут дать основу для глубокого анализа того, почему и каким образом человек приходит к решению о самоубийстве.

Отметим наличие взаимозависимости: с одной стороны, выбор алгоритма, равно как и интерпретация результатов его использования, зависят от идей, заложенных в выборе понятий и их операционализации; с другой стороны, способ операционализации в значительной мере определяется тем, как мы априори видим алгоритм анализа данных, как собираемся интерпретировать результаты его применения (эта часть нашего утверждения менее традиционна; ее подробное обоснование см. в [Толстова, 1998а]; см. также примеры, приведенные ниже в настоящем параграфе и в разделе 5). Связь выбранного алгоритма с тем, как мы будем интерпретировать найденную формальную закономерность, представляется очевидной. Другими словами, три этапа – (1) измерение, (2) выбор и реализация конкретного алгоритма анализа и (3) интерпретация получающихся результатов неразрывно связаны друг с другом. То, каким способом реализуется один из них, обусловливает способы реализации двух других.

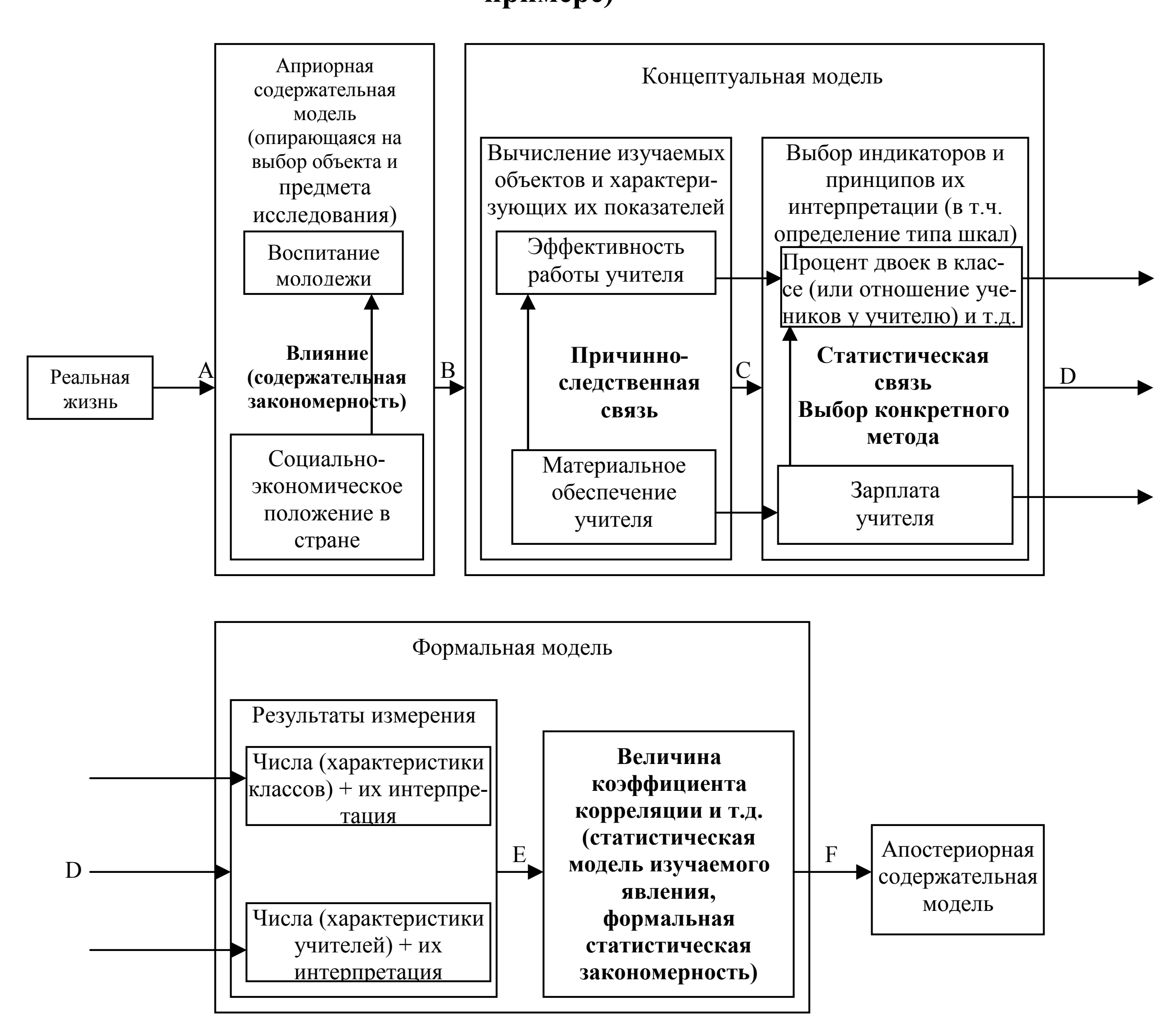
Итак, реализация алгоритма приводит нас к искомой формальной (математической) модели изучаемой социальной реальности.[16] Интерпретация этой модели позволяет сделать содержательные выводы, т.е. фактически приводит исследователя к *апостериорной содержательной модели* той же реальности. Подчеркнем принципиальное отличие этой модели от того, что выше мы назвали априорной содержательной моделью. Апостериорная содержательная модель "вбирает в себя" все модельные свойства описанных выше априорной содержательной, концептуальной и формальной моделей. Если неадекватными реальности были наши априорные содержательные представления о ней, измерение, выбор метода и интерпретация результатов его применения, то такой же неудачной будет и наша итоговая содержательная модель.

В социологии острота проблемы адекватного соотнесения реальности с ее формальной (математической) моделью объясняется, в первую очередь, тем, что построение и априорной содержательной, и концептуальной моделей в значительной мере определяется субъективным видением мира социологом. В частности, здесь мы никуда не уйдем от известного веберовского принципа отнесения к ценности. Кроме того, практически каждое социологическое явление даже при тщательной отработке априорной содержательной и концептуальной моделей оказывается возможным формализовать многими способами Для решения одной и той же задачи, как правило, существует несколько методов, приводящих, вообще говоря, к разным выводам. Это положение представляется нам принципиальным (по крайней мере, для современного состояния науки): интересующие социолога явления столь сложны и многогранны, что любая формализация приводит к учету лишь какой-то стороны каждого

явления, разные методы отвечают разным сторонам. Чтобы преодолеть соответствующие трудности, можно использовать специальные подходы (в первую очередь - комплексное использование нескольких методов [Толстова,1991а]).

Поясним рассмотренные положения на условном, заведомо упрощающем и реальность, и подходы к ее изучению, примере (см. схему 1).

**Формирование и операционализация понятий при анализе данных (на условном примере)***




* Жирным шрифтом выделены блоки, касающиеся процесса формализации понятия закономерности. Внутренние прямоугольники отвечают блокам, касающимся процесса измерения. Не отмечены многочисленные обратные связи.

Раскрытие связок: А – абстрагирование от реальности на основе взглядов исследователя, формирование представлений об объекте и предмете исследования, выделение основных понятий и связывающих их закономерностей через отнесение к ценности; В – концептуализация:

формирование ЭС и МС (см. п.2.2), формирование и операционализация понятий с учетом "взаимодействия" исследователя и респондента; C – операционализация понятий; D – определение измеряемых объектов (построение и корректировка выборки), непосредственная реализация процедуры измерения; E – реализация метода анализа данных; F – интерпретация результатов применения метода

Предположим, что мы хотим изучить влияние социально-экономического положения в стране на воспитание молодежи. Сначала – об априорной содержательной модели. Сама постановка задачи говорит о том, что по существу мы уже опираемся на какие-то априорные модельные соображения, когда формулируем проблему именно указанным образом: кто-то, может быть, не согласится с использованием выражения "социально-экономическое положение в стране" – дескать, не о чем тут говорить – обычное положение, типичное для стран, "строящих" капитализм (для сравнения заметим, что вряд ли сам термин "социально-экономическое положение в стране" мог бы фигурировать в постановке задач советскими социологами 20 лет назад); кто-то не согласится с тем, что рассматривается именно проблема воспитания молодежи – дескать, более актуальным является анализ положения мелкого предпринимателя и т.д.

О реальных объектах во всей их уникальности и содержательном многоцветье, мы пока имеем смутное представление: это предположительно либо молодежь, либо дети, либо те, кто их в том или ином смысле воспитывает (воспитатели детских садов, учителя, деятели культуры, средств массовой информации и т.д.). Именно в их характеристиках (пока нам неизвестных) так или иначе проявляется и социально-экономическое положение, и проблемы воспитания. Об отношениях между реальными объектами, условно названных нами содержательной закономерностью, тоже пока известно мало; мы просто предполагаем, что социально-экономическое положение как-то влияет на воспитание молодежи.

Перейдем к обсуждению концептуальной модели. Будем рассматривать только учителей (тем самым вычленим изучаемые объекты), выявим, как наша проблема проявляется в их жизни. Выделим некоторые стороны жизни реальных учителей, относительно которых скажем только то, что они адекватно отражаются понятиями "материальное положение учителя" и "производительность его труда" (именно здесь речь идет о рождении понятий, отражающих качества изучаемых объектов, о формировании показателей). Будем полагать, что нас интересует причинно-следственное отношение между упомянутыми содержательными аспектами жизни учителя (первый шаг в концептуализации изучаемой закономерности). Ясно,

что и на этом этапе мы использовали наше субъективное представление о проблеме. Кто-то, может быть, сочтет нужным придти к другим понятиям – скажем, не к “материальному положению”, а к “психологическому диссонансу, возникающему у учителя вследствие резкой смены ценностных ориентаций, господствующих в обществе”.

Чтобы завершить построение концептуальной измерительной модели, мы должны найти способ выражения названных понятий через наблюдаемые признаки, т.е. осуществить их операционализацию. Скажем, считаем, что первое понятие хорошо отражается признаком "зарплата учителя", а второе – признаком "средний процент успеваемости в тех классах, где учитель работает". Здесь тоже мы субъективны. Кто-то, может быть, оспорит предложенную операционализацию: скажем, будет считать, что о материальном положении лучше судить по тому, куда учитель посылает отдыхать собственных детей, а об эффективности работы учителя – скажем, по числу учеников, говорящих о своей любви к данному учителю (а выявление подобных аспектов взаимоотношений учителя и ученика может потребовать использования неформализованных методов опроса) и т.д. Но, так или иначе, будем считать, что, выбрав указанные выше признаки, мы тем самым построили концептуальную модель измерения.

Для завершения построения концептуальной модели искомой содержательной закономерности необходимо выбрать конкретный способ измерения связи между упомянутыми признаками – тот, который мы считаем хорошим отражением интересующей нас причинно-следственной зависимости. Например, в качестве меры связи может служить обычный коэффициент корреляции Пирсона.

Выбрав конкретную меру, мы как бы операционализируем интуитивное представление о связи и тем самым полностью концептуализируем модель искомой закономерности. Вычислив конкретное значение этой меры (например, получив значение коэффициента корреляции, равное 0,8), мы тем самым получаем интересующую нас формальную закономерность, формальную модель изучаемого явления.

Но здесь снова нет однозначности, снова много субъективности.

Чтобы подробнее пояснить это, подчеркнем наличие необходимости “разводить” две компоненты описанной модели изучаемого явления: первую – выбор в качестве меры связи именно такого показателя связи, а не другого; и вторую – конкретное значение выбранного коэффициента именно для анализируемой системы. Мы обращаем на это внимание читателя, поскольку на практике исследователи-социологи иногда учитывают только вторую компоненту. Вопрос же о выборе первой даже не ставится: считается само собой разумеющимся, что используется наиболее употребительный метод соответствующего плана. Это вряд ли может

быть оправдано.

Как мы отмечали, сложность формализации социальных явлений приводит к тому, что для решения практически любой социологической задачи существует много методов. Это в полной мере отвечает нашей ситуации. В статистике известно более сотни способов измерения показателей связи между двумя признаками. Каждый из них отражает лишь какую-то одну сторону "истинной" связи. Выбрав показатель, мы тем самым волей-неволей выбираем соответствующий "срез с реальности". И у нас всегда останется вопрос о том, отвечает ли используемый формализм нашим содержательным представлениям о сути изучаемого явления. Например, о том, можем ли мы считать, что, скажем, обычный (Пирсоновский) коэффициент корреляции хорошо отражает природу формирования у учителя настроя на эффективную работу? Или же надо использовать, скажем, ранговый коэффициент Кендалла, или какой-либо из энтропийных коэффициентов связи? Вероятно, лишь имея перед собой множество подобных коэффициентов, мы можем понять, чт**о** есть наша связь в реальности (подробнее о сути различных подходов к измерению связи между двумя номинальными признаками можно прочесть, например, в [Лакутин О. В., Толстова Ю.Н., 1990, 1992]).

Теперь – несколько слов о том, как в нашем примере может проявляться связь между измерением, выбором алгоритма и интерпретацией результатов реализации последнего. Наиболее очевидной представляется связь между типом шкал (определение которого является этапом процесса измерения) и используемым коэффициентом связи. Так, известно, что существуют коэффициенты связи, рассчитанные на номинальные шкалы (например, коэффициенты, основанные на известном критерии Хи-квадрат, о них пойдет речь во второй части книги); порядковые шкалы (например, известные коэффициенты Спирмена и Кендалла), интервальные шкалы (например, классический коэффициент Пирсона). Тип же используемых шкал определяется многими обстоятельствами, в том числе содержательной интерпретацией используемых при измерении чисел. Предположим, например, что материальное положение учителя измеряется его зарплатой и обсудим, каким можно будет считать тип используемых шкал при разных содержательных посылках.

Соответствующую шкалу можно будет считать дихотомической номинальной, если, скажем, мы поделим всех учителей на тех, которые получают зарплату, не превышающую нижнюю границу уровня бедности (т.е. не превышающую стоимость т.н. потребительской корзины), и тех, зарплата которых превышает эту границу. При определенных условиях ту же шкалу можно считать порядковой – скажем, выделять три группы учителей – (1) не обеспеченных даже на уровне потребительской корзины (т.е., по официальной терминологии,

живущих в нищете), (2) имеющих возможность оплатить потребительскую корзину, но не более того (живущих в бедности) и (3) обеспеченных хотя бы насколько-то выше этого уровня. А возможно полагать, что мы имеем дело с интервальной шкалой – считать, что, скажем, различие между учителями, получающими 3400 и 3600 рублей с интересующей нас точки зрения – та же, что и между учителями, получающими 400 и 600 рублей (мы намеренно описываем ситуации, когда тип шкалы определяется не технологией получения шкальных значений, а содержательными рассуждениями исследователя; подробнее об этом можно прочесть в [Толстова, 1998]). Ясно, что в каждом из описанных случаев мы должны выбрать свой, подходящий для используемой шкалы, коэффициент корреляции.

Более тонким является вопрос о связи измерения и выбора алгоритма с тем, как мы собираемся интерпретировать результаты применения последнего Конечно, прежде всего в интерпретацию результатов использования какого-либо алгоритма вкладывается тот смысл, который является естественным для метода – скажем, значение коэффициента корреляции мы будем интерпретировать как связь между некоторыми явлениями и, вероятно, при этом будем пытаться "выйти" на какие-то причинно-следственные отношения. Но в рамках такого "естественного" подхода имеются нюансы (подробнее о "естественном" подходе к интерпретации результатов анализа данных и необходимости расширения такого подхода см. [Толстова, 1991а]).

Применительно к рассматриваемому примеру отметим лишь то, что приведенные выше рассуждения об определении типов шкал напрямую касались выбора предполагаемых способов интерпретации результатов измерения связи. Поясним это.

Выбор обычного коэффициента корреляции Пирсона означает наше желание того, чтобы к выводу о наличии связи мы приходили, скажем, наблюдая, что при переходе зарплаты от 400 к 600 рублям эффективность работы учителя в среднем возросла на столько же, насколько в среднем она возросла при переходе от 3400 к 3600 рублям. Другими словами, если такого рода соотношения действительно имеют место (по всем парам значений наблюдаемых признаков), мы получим коэффициент, близкий к единице, и будем делать вывод о наличии соответствующей содержательной связи. И, рассуждая подобным образом, рискуем уйти в сторону от действительного положения вещей. Повышение зарплаты учителя от 3400 до 3600 рублей действительно можно интерпретировать как получение учителем возможности, скажем, регулярно покупать новые книги и, вследствие этого, более эффективно работать с учениками. А вот повышение зарплаты от 400 до 600 рублей вряд ли правомерно интерпретировать таким

же образом. И та, и другая зарплата не могут обеспечить учителя даже возможностью наполнить продовольственную (а не то что потребительскую!) корзину. Причины же имеющего место фактически более высокого качества работы учителей, получающих зарплату 600 рублей (по сравнению с учителями, получающими 400 рублей), вероятно, надо искать в чем-то другом. Скажем, может оказаться, что 600 рублей получают учителя с более высоким стажем, более солидного возраста, и, вследствие этого, получившие "классическое" советское воспитание в духе осознания необходимости соблюдения долга, любви к ближнему, привыкшие работать "за идею" и т.д. Ясно, что здесь вывод о том, что улучшение материального положения способствует повышению качества работы учителя – ни при чем. А вот если мы будем использовать какой-либо из упомянутых выше порядковых коэффициентов корреляции, возможности интерпретации будут другие. Так, если окажется, что люди, живущие в нищете, в среднем хуже работают, чем люди, живущие в бедности, а последние – в среднем хуже, чем те, которые смогли "вылезти" из бедности, наверное, у нас будут основания говорить о подтверждении нашей закономерности. Нетрудно видеть, что именно такой вывод позволит сделать близость к единице какого-либо из порядковых коэффициентов.

Подведем итог рассмотрению нашего примера.

Совокупность измеренных ("наблюденных") значений выбранных признаков (т.е. найденные для каждого учителя значения его зарплаты и среднего процента успеваемости в его классах) – это наши исходные данные (формальные), это первичная, полученная в результате измерения, математическая модель некой реальности – той, которую мы считаем нужным отобразить при измерении в математические конструкты. Рассчитанное нами значение выбранного формального показателя связи между упомянутыми признаками – формальная закономерность, математическая модель изучаемого явления.

Однако процесс осуществляемого с помощью анализа данных моделирования реальности включает в себя не только применение конкретного алгоритма (расчет коэффициента корреляции Пирсона), но и (1) все, что этому расчету предшествовало: и само формирование в нашем мозгу представлений об исходных понятиях вкупе с гипотезами о наличии связи между ними, и операционализацию этих понятий, и выбор именно такого-то алгоритма, а не другого и т.д., а также (2) все, что должно следовать за указанным расчетом, за реализацией конкретного алгоритма, т.е. интерпретацию полученных результатов. Другими словами, говоря о статистической (и, вообще, - математической) закономерности как о модели реальности, сопрягая понятие такой закономерности, в первую очередь, с тем, что заложено в используемом алгоритме анализа данных, мы в то же время будем иметь в виду не только собственно

реализацию алгоритма, но и то, что ее предваряет, и то, что за ней следует. См. об этом также раздел 5.

Ниже содержательную закономерность, отвечающую априорной содержательной модели, мы условно будем называть *социологическим явлением* (понимая его не как элемент диады "сущность - явление", а чисто интуитивно - как нечто, происходящее в обществе). Формальную закономерность (часть формальной модели) будем называть просто *закономерностью* (или статистической закономерностью, поскольку в данной работе нас интересуют именно соответствующие соотношения). Совокупность формальных данных будем называть эмпирическими фактами. Под формальными данными при этом будем иметь в виду или результаты измерения, или (хотя и реже) – результаты работы математического алгоритма (полагаем, что, например, утверждение, состоящее в том, что коэффициент корреляции равен, скажем, 0,8, вполне считать эмпирическим фактом).

В заключение настоящего параграфа отметим, что мы рассмотрели, конечно, не все аспекты проблемы соотнесения формального и содержательного при изучении статистических социологических закономерностей. Так, мы не рассмотрели ситуацию, когда у социолога априори, до проведения исследования, не сформированы (вообще, или в достаточной мере) представления о содержательной закономерности, когда, реализуя этап измерения, исследователь действует в значительной мере по наитию, и в процессе анализа собранных данных как бы “прощупывет” их с помощью многих методов, выбирая затем те результаты, которые согласуются с его знаниями (как априорными, так и полученными в результате сравнительного анализа результатов применения, вообше говоря, разных методов). Почти не коснулись мы и того, что нетривиальные причинно-следственные отношения, вероятно, могут быть обнаружены лишь с помощью творческого использования т.н. качественных методов, что применение последних, вероятно, необходимо и при формировании инструмента измерения и т.д. и т.д.

## 1.4. Статистическая закономерность как результат "сжатия" исходных данных

Посмотрим на проблему поиска статистических закономерностей с иной точки зрения. Поставленный в п. 1.1 вопрос о том, как "увидеть" в матрице "объект-признак" интересующие нас закономерности, можно сформулировать по-другому: как сжать исходную информацию,

чтобы искомые закономерности предстали перед нами в явном виде? Известные способы сжатия – это и суть алгоритмы анализа данных. Поясним более подробно, какое "сжатие" здесь имеется в виду.

Начнем с того, что любая выявленная в процессе научного исследования закономерность (и не только в социологии) является определенного рода сжатием какой-то информации об изучаемых объектах, имеющейся в распоряжении исследователя. Виды такого сжатия весьма разнообразны. Выбор конкретного вида зависит от исследователя и определяется его априорными представлениями о характере изучаемого явления, пониманием цели сжатия. Коснемся двух аспектов таких представлений.

Первый аспект касается формальной сути алгоритмов сжатия. Дело в том, что в интересующем нас случае (когда рассматриваются только статистические закономерности) результаты такого сжатия чаще всего выражаются в виде определенных характеристик частотных (вероятностных) распределений значений рассматриваемых признаков (подробнее об этом пойдет речь в разделе 3). Так, совокупность из 1000 значений какого-либо признака может быть сжата до одного числа - соответствующего среднего арифметического значения. Множество из 2000 значений двух признаков можно сжать до одного числа – какого-либо коэффициента парной связи между этими признаками. Совокупность из 10000 значений 10-ти признаков может быть сжата до 9-ти коэффициентов регрессионного уравнения, связывающего один из рассматриваемых признаков с 9-ю другими и т.д.

Второй интересующий нас аспект представлений исследователя, выбирающего алгоритм анализа данных, касается некоторых моментов трактовки роли сжатия исходной информации в выявлении любых интересующих человека закономерностей природы (общества). Мы имеем в виду определенные стороны понимания самого термина "закономерность". Здесь, в свою очередь, выделим два момента.

Во-первых, при выявлении любой закономерности практически всегда неизбежна потеря исходной информации об изучаемых объектах (здесь мы не говорим о том, что эта потеря может быть не "абсолютна", все исходные данные могут быть сохранены, скажем, где-то в памяти ЭВМ): такова "цена" найденных исследователем научных положений. Казалось бы, это утверждение довольно естественно. С потерей информации тот же социолог сталкивается на каждом шагу. Скажем, она происходит уже благодаря использованию анкетного опроса (т.е. при сборе данных, еще до всякого анализа; здесь представляется уместным отметить, что, в соответствии со сказанным в предыдущем параграфе, в социологии отсутствует четкая граница между сбором и анализом данных), в таком случае вместо живого, неповторимого человека мы

имеем набор чисел - ответов этого человека на вопросы анкеты. И необходимо тщательно "отслеживать", правомерны ли допускаемые потери (в частности, надо решить поставленные выше вопросы: те ли признаки мы выбрали для характеристики интересующих нас процессов, так ли определили тип шкалы, правильно ли заранее спрогнозировали, какой смысл будем вкладывать в числа, получающиеся в результате реализации алгоритма анализа данных и т.д.).

Подобные вопросы очень актуальны для социологии. Процесс поиска ответов на них далеко не всегда прост. Но суть соответствующих процедур в значительной мере состоит в выявлении того, какую информацию мы можем позволить себе потерять при сборе и анализе данных.

Во-вторых, во многих алгоритмах анализа встает вопрос о степени сжатия исходной информации. Например, в агломеративных алгоритмах классификации (т.е. таких, в соответствии с которыми разбиение совокупности на классы осуществляется в процессе реализации целой серии шагов, на первом из которых каждый исходный объект являет собой отдельный класс, а на последнем - все объекты объединяются в единый класс; описание подобных алгоритмов можно найти, например, в книге [Статистические методы …, 1979. Гл.12]; заметим, что именно агломеративные алгоритмы классификации заложены в известном пакете программ SPSS) встает вопрос, какое разбиение выбрать, сколько классов это разбиение должно содержать. В алгоритмах многомерного шкалирования (или, например, факторного анализа) требуется ответить на вопрос о том, какова размерность искомого признакового пространства, т.е. сколько латентных переменных определяют интересующее нас явление и т.д.

Наиболее естественным ответом на подобные вопросы, вероятно, можно считать тот, в соответствии с которым сжатие должно осуществляться до тех пор, пока человеческий разум не окажется способным охватить единым взглядом полученный результат. Иначе то, что формально могло бы вроде считаться закономерностью, для нас фактически таковой не будет. Так, строя типологию каких-либо объектов с помощью упомянутых методов классификации, мы при любой постановке задачи вряд ли сможем разумно проинтерпретировать как типологию, скажем, совокупность из 200 классов, каждый из которых характеризуется 15 признаками. В подобной ситуации мы, вероятно, поставим перед собой задачу дальнейшего сжатия исходной информации. То же можно сказать и о той ситуации, когда мы выявили 200 латентных факторов, формирующих пространство восприятия респондента, найденное с помощью многомерного шкалирования. Оси 200-мерного пространства мы даже и не будем называть латентными факторами[17].

Заметим, что рассмотренные аспекты понимания искомой закономерности касаются

одного из аспектов проблемы интерпретации результатов применения математического метода.

## 1.5. Основные цели анализа данных

Итак, в соответствии со сказанным выше, основная цель анализа данных - выявление (подтверждение, корректировка) каких-то интересующих исследователя статистических закономерностей; или, другими словами, - определенного рода сжатие, усреднение содержащейся в данных информации. Однако мы не можем ограничиться только такой формулировкой. Она нам говорит лишь о формальной стороне действий социолога, изучающего эмпирические данные. Но естественно, что за выбором алгоритма анализа не могут не стоять содержательные соображения, о чем мы частично уже говорили. Причины, побуждающие исследователя искать ту или иную закономерность, могут быть разными. Это должно учитываться в процессе анализа.

Ниже мы коротко рассмотрим те стороны априорных концепций ученого, которые должны играть роль при определении общей стратегии работы. Речь пойдет о вопросах, обычно относимых к области функций научного исследования. Эти вопросы серьезны и не достаточно основательно разработаны применительно именно к социологии. Будучи ограниченными жанром настоящей работы, мы будем "скользить по поверхности". Однако хотелось бы, чтобы читатель почувствовал скрывающуюся под этой "поверхностью" глубину. Изучению того, каковы функции научного исследования, уделяли огромное внимание такие крупные ученые, как О.Конт, Дж.С.Милль, Э.Мах, К.Поппер, К.Гемпель и другие. Много работ соответствующего плана имеется и в отечественной литературе. Для интересующегося читателя мы назовем лишь выпущенные в последние годы учебные пособия [Степин и др., 1995; Философия и методология науки, 1996], и ставшие классическими работы [Лакатос, 1995; Поппер, 1983]; см. также [Ядов, 1998, с.53-62].

Задачу поиска закономерности иногда отождествляют с задачей *объяснения* интересующего исследователя явления (напомним, что главный смысл объяснения состоит в подведении объясняемого явления под какой-либо закон, см. также [Девятко, 1996; Терборн, 1994]; подчеркнем, что здесь явление – это не обязательно наша содержательная закономерность; см. об этом ниже). Конечно, достижение соответствующей цели (точнее, реализации отвечающей ей функции науки) является актуальной в любом социологическом исследовании. Вероятно, ее почти всегда можно считать основной целью анализа. Так, выяснив

в приведенном в п. 1.3 примере, что коэффициент корреляции между уровнем безработицы и числом суицидов в регионе близок к единице, мы считаем, что самоубийство объясняется материальной неустроенностью человека. Однако этот же пример показывает сложность процесса объяснения. Поясним это.

Упомянутая сложность снова начинается с понимания используемых терминов. То явление, которое мы объясняем, можно понимать по-разному. Во-первых, его можно отождествить с совокупностью наблюдаемых фактов (т.е с формальными данными в нашей терминологии). В рассматриваемом примере – это пары значений уровня безработицы и частоты суицидов в регионах. Тогда закон, под который мы "подводим" явление – это и есть найденный коэффициент корреляции. Величина коэффициента говорит о наличии статистической связи, что как бы объясняет, почему в наблюдаемых данных б**о**льшим значениям уровня безработицы отвечают б**о**льшие частоты суицидов (потому, что между соответствующими признаками имеется сильная статистическая связь). Здесь представляется уместным вспомнить, что статистическая связь, вообще говоря, не доказывает наличие причинно-следственной (см. сноску 15). Выявление статистической закономерности - это как бы формальное объяснение того, что в действительности интересует социолога. Хотя такое объяснение зачастую играет огромную роль в исследовании, социолог, как правило, стремится им не ограничиваться. Вероятно, с объяснением можно отождествлять выявление причинно-следственных отношений. А это чаще всего бывает возможно сделать как мы отмечали в конце п. 1.3, только на основе применения качественных методов.

Во-вторых, объясняемое явление можно понимать так, как мы предложили это делать выше (в конце п.1.3) – как содержательную закономерность в нашем смысле. Для рассматриваемого примера – это содержательные представления о том, что невозможность найти работу подталкивает человека к самоубийству. В таком случае расчет упомянутого выше коэффициента корреляции можно рассматривать как формальную закономерность, отвечающую этой содержательной закономерности и подтверждающую ее. Тогда "закон", под который мы подводим объясняемое явление, можно отождествлять с теми самыми причинно-следственными отношениями, о котором шла речь выше.

Только поиском объяснения цели научного исследования обычно не ограничиваются. Наряду с объяснением изучаемого явления, представляется целесообразным всегда иметь в виду по крайней мере еще две цели: описание исходных данных и осуществляемое на основе выявленной закономерности предсказание того или иного явления. *Описание* - цель, достичь которую часто бывает необходимо прежде, чем непосредственно приступать к поиску основной

интересующей исследователя закономерности (однако некоторые ученые - например, Э.Мах - полагали, что описание – единственная функция научного исследования; объяснение и предвидение, по Маху, сводятся к описанию). Предсказание тоже зачастую считается основной целью научного исследования (ср. с известным афоризмом О.Конта: "Знать, чтобы предвидеть"), и с этим трудно спорить.

Описание требуется для того, чтобы исследователь мог хотя бы самым приблизительным образом сориентироваться в том "море" данных, о котором шла речь выше. А потребность в этом обычно имеется. Ведь далеко не всегда социологу бывает с самого начала полностью ясно, каков вид закономерностей, "скрывающихся" за интересующими его данными, не всегда понятно, например, какими признаками эти закономерности должны описываться, возможно ли в принципе подобрать соответствующие признаки и т.д. Описание может помочь дать ответы на подобные вопросы с тем, чтобы потом уже можно было более направленно искать интересующие социолога соотношения. Описание данных обычно достигается с помощью самых простых способов сжатия исходных данных. Примеры: доля женщин в изучаемой совокупности; средний возраст респондентов; величина разброса респондентов по возрасту (например, выраженная в виде соответствующей дисперсии); наиболее часто встречающаяся среди респондентов профессия; нижний уровень дохода 10 % самых богатых респондентов и т.д. Заметим, что совокупность наиболее употребительных приемов получения закономерностей, описывающих изучаемое множество объектов, называется *описательной, или дескриптивной, статистикой*. Это – одна из областей анализа данных (см. раздел 1 части 2).

Прогноз тех или иных характеристик жизни общества по существу служит целью выявления любой закономерности: изучать ту или иную сторону жизни общества чаще всего надо для того, чтобы научиться управлять какими-либо процессами. Прогноз осуществляется обычно с помощью довольно сложных алгоритмов. Часто методы анализа данных (в качестве "прогнозных" методов могут использоваться, например, алгоритмы регрессионного анализа, см. п. 2.6.2 части II) здесь сопровождаются полуформализованными процедурами построения экспертных сценариев (см., например, [Задорин, 1994]) .

Для понимания сути анализа данных важно отметить, что и при описании данных, и при прогнозе могут использоваться алгоритмы того же рода, что и при поиске основной закономерности. Границы между этими тремя целями часто бывают размыты. Кроме того, можно выделить и другие цели[18]. Упомянем здесь лишь одну из них, лежащую в русле уже упомянутой нами гуманитарной парадигмы – *понимание* изучаемого явления.

Как известно, термин "понимание" как название одной из главных функций науки с

конца XIX века является ключевым для социологии. Если творчество О.Конта было шагом вперед в том смысле, что он одним из первых сказал, что социология – такое же строгое направление в науке, как и ее естественные ветви, и был явным сторонником того, что в наше время называют социологическим реализмом (мы полагаем, что это было шагом вперед, хотя в современной отечественной литературе принято "ругать" Конта за то, что он, говоря о методах социологии, "не усмотрел" человека; на наш взгляд, подобная "критика" не учитывает исторических условий жизни основоположника социологии), то к названному периоду стала ясна необходимость обращать больше внимания на мотивы поведения отдельных людей, т.е. необходимость учета постулатов социологического номинализма. В творчестве В.Дильтея родился термин "понимающая психология", в творчестве М.Вебера – термин "понимающая социология" (красноречиво выглядит также то, что В.А.Ядов при последнем переиздании своей известной книги по методике социологических исследований [Ядов, 1998] снабдил ее подзаголовком: "описание, объяснение, понимание социальной реальности").

В литературе уделяется огромное внимание анализу сходства и различия смыслов терминов "объяснение" и "понимание" как отражений соответствующих функций науки. Как известно, с именем Дильтея связано разделение наук на науки о природе и науки о духе (социология принадлежит к числу последних). Бытует точка зрения, в соответствии с которой главная познавательная функция наук о природе – объяснение (подведение единичного объекта под общий закон, в результате чего уничтожается неповторимость объекта), а наук о духе – понимание (т.е. изучение объекта в его неповторимости). Мы присоединяемся к другому мнению, в соответствии с которым любая наука (это особенно касается наук о человеке и, в частности, социологии) должна и объяснять, и понимать (свое "понимание" имеется, скажем, даже в математике; этого мы здесь не касаемся) .

Мы не можем не упомянуть о понимании как об одной из познавательных функций социологии в силу огромной важности достижения понимания изучаемого объекта (человека) в любом социологическом исследовании. Однако, поскольку в данной работе нас интересует только анализ данных, то ограничимся сказанным и напоминанием читателю того, что "понимание" обычно достигается с помощью мягких методов исследования, что для анализа их результатов существует масса приемов, составляющих т.н. анализ качественных данных, о котором мы уже говорили в п. 1.2. Вернемся к описанию, объяснению, предсказанию.

Подчеркнем, что выше мы везде неявно предполагали, что для описания какого-либо явления, выявления определяющих его причин, предсказания последствий и т.д. необходимо использование математики. Мы считали очевидным, само собой разумеющимся, что

анализировать данные, изучать на этой основе окружающую нас реальность, можно только с помощью математических методов. А так ли это? Этот вопрос тем более актуален, что любому социологу не раз приходилось слышать о том, что использование математики в социологии связано с определенными проблемами.

Теперь попытаемся пояснить, почему процесс анализа данных должен опираться на применение математического аппарата, и какого рода сложности возникают при использовании математики в науке вообще и в социологии в частности.

# 2. МАТЕМАТИЧЕСКИЕ МЕТОДЫ КАК СРЕДСТВО ПОЗНАНИЯ СОЦИАЛЬНЫХ ЯВЛЕНИЙ

## 2.1. Роль математизации научного знания

О роли математизации научного знания в литературе говорится довольно много. Ниже мы коснемся лишь некоторых аспектов соответствующей проблемы, играющих, на наш взгляд, серьезную роль в организации эмпирического социологического исследования.

К сожалению, в кругу социологов часто бытует мнение о том, что математические методы как бы противостоят настоящей "гуманистической" социологии. И в этом смысле термин "математический" отождествляется с термином "количественный", понимаемым в приведенном выше смысле (имеем в виду известную пару "количественная социология" – "качественная социология"). А это, с нашей точки зрения, – кардинально неправильное положение, мешающее эффективному развитию социологии. Поясним это.

Везде, где мы хотим говорить о науке, требуется определенный уровень четкости, конструктивности рассматриваемых положений. Никакой "поток сознания", исходящий от респондента, взгляды которого изучаются "мягкими" методами, не позволит нам говорить о научных выводах, если мы в этом "потоке" не выделили некоторые "жесткие" логические конструкты (заметим, что, проанализировав подходы, используемые авторами качественных исследований, можно придти к выводу, что научный характер эти разработки приобретают, благодаря отображению каждого респондента в некую лингвистическую полуформализованную структуру и получение теоретических выводов на базе определенной агрегации подобных структур, полученных от разных респондентов, принадлежащих к одной субкультуре), а определенный уровень четкости этих конструктов позволяет говорить об использовании

математического языка. Математика, собственно говоря, начинается везде, где нам удается достаточно четким образом обрисовать интересующую нас жизненную ситуацию. Например, математический аппарат можно использовать уже на стадии изучения формирования понятия в сознании респондентов [Толстова, 1996б]. Другой пример – сами "качественники" говорят о том, что последовательная реализация соответствующего подхода приводит к рождению понятия переменной [Семенова, 1998. С. 198], а это – уже вполне "количественная" ситуация, делающая естественным использование математического аппарата. Да и в типично качественном исследовании, как мы уже отмечали (и как свидетельствует приведенный в конце работы список публикаций) математика активно используется.

Иногда при серьезном изучении мнений респондента удается использовать традиционный математический язык и соответствующие математические теории. Скажем, плюралистичность мнения одного респондента (т.е. то обстоятельство, что он в разное время, при разных условиях, вообще говоря, будет по-разному оценивать один и тот же объект), о которой мы говорили в п.1.2 может быть четко описана фразой: мнение одного респондента об одном объекте имеет нормальное распределение. Именно это описание было использовано в упомянутой там терстоуновской модели метода парных сравнений. О том, какую роль математический язык играл в творчестве известных специалистов в области социологического измерения, говорится в [Толстова, 1998].

Бывают ситуации, когда известные математические теории “не работают”, и тогда рождается новая ветвь математики. Так родились теория измерений, многомерное шкалирование. Известный американский социолог П.Ф.Лазарсфельд, глубоко анализируя соотношение наблюдаемого и ненаблюдаемого (ответов респондентов на вопросы анкеты и скрытых факторов, определяющих эти ответы), разработал соответствующую теорию, сформулированную им на математическом языке и названную латентно-структурным анализом. Отметим также попытки создания математической социологии такими известными исследователями, как Г.М. Блейлок, предложивший базирующуюся на причинном анализе теорию конфликта [Blalock, 1989]; Дж.Коулмен, создавший теорию анализа временных рядов, математическую теорию коллективного действия [Coleman, 1990].

Ярким примером математизации сугубо "гуманитарных" взглядов исследователя является детерминационный анализ, автор которого отверг традиционные подходы к шкалированию, счел неадекватным сути социологии использование в ней математической статистики, но, тем не менее, пришел к математике – правда, весьма своеобразной, являющейся

обобщением аристотелевской силлогистики [Чесноков, 1982, 1985].

Для творчества каждого названного ученого характерно то, что, по большому счету, в нем речь идет о предложении определенного языка, адекватно описывающего рассматриваемые социальные процессы.

Однако, конечно, следует отметить, что в наше время вряд ли мы можем говорить о состоятельности претензии кого бы то ни было на разработку математической социологии. Пока соответствующей формализации поддались лишь сравнительно простые социологические образования. Математика всегда отражает все же относительно простые ситуации (здесь представляется целесообразным вспомнить Конта, который в своей классификации наук математику отнес к самым простым, а социологию – к самым сложным [Конт, 1996]). Правда, это говорит еще и о наличии серьезных проблем с разработкой социологических теорий.

## 2.2. Априорная модель изучаемого явления.

### Эмпирическая и математическая системы.

Чтобы прочувствовать специфику использования математических методов как средства познания социальных явлений, взглянем на отношение математики к реальности с несколько иных позиций, чем это было сделано выше. То, о чем пойдет речь, как бы лежит “за кадром” всего сказанного ранее в разделе 1.

Возможность применения математики возникает тогда, когда исследователь абстрагируется от многих конкретных черт изучаемого объекта и предполагает адекватной сути решаемой задачи определенную формализацию рассматриваемого явления. Подчеркнем последний момент. Речь идет о том, что априори, т.е. прежде, чем осуществлять какой бы то ни было математический анализ данных (и даже прежде, чем получать эти данные), необходимо сформировать определенное представление о том, каков характер подлежащего изучению явления (эти представления лежат в основе того, что в п.1.3 названо априорной содержательной и концептуальной моделями). Совокупность таких представлений можно назвать *априорной моделью этого явления*, должны быть достаточны для того, чтобы на их основе можно было выбрать (разработать) и способы сбора данных, и подходы к их интерпретации, и формальный аппарат для непосредственного анализа данных, и принципы интерпретации результатов применения этого аппарата. И роль социолога при формировании описанной априорной модели является главной (по сравнению с ролью математика).

Переходя к более подробному логическому анализу рассматриваемого процесса, можно сказать следующее. Применение математики опирается на то, что мы считаем возможным (1) выделить некоторый фрагмент реальности; (2) построить (посредством измерения) его математическую модель (т.е. получить исходные данные); (3) изучить эту модель традиционными для математики способами (в нашем случае - применить тот или иной алгоритм анализа данных) и прийти к некоторым выводам о ее "устройстве" (в результате анализа данных получить какой-то математический результат: вычислить точное значение коэффициента корреляции, найти параметры уравнения регрессии и т.д.); (4) проинтерпретировать эти выводы на содержательном языке (т.е., как говорят обычно, проинтерпретировать результаты анализа данных) и получить таким образом новое знание о реальности. Первые два этапа обычно относят к области измерения (шкалирования), последние два - к области собственно анализа данных. Но все четыре этапа тесно связаны друг с другом, их нельзя рассматривать по отдельности. Реализация этих этапов приводит к построению сложной модели реальности, первым шагом которого является построение некоторой первичной модели – результата измерения. Соответствующий процесс обычно бывает связан с решением ряда не всегда простых (особенно для социологии, поскольку она имеет дело с весьма сложной реальностью) проблем. Рассмотрим формальную сторону этого процесса более подробно.

Строя первичную модель в процессе измерения, т.е. реализуя первые два этапа, мы должны вычлененить круг рассматриваемых объектов; ограничить множество их свойств лишь теми, которые интересуют исследователя; вычленить те отношения между объектами (рассматриваемыми как носители выделенных свойств), которые должны моделироваться в процессе измерения. (В п. 1.3 мы по существу с несколько иной точки зрения рассматривали тот же процесс, говоря о рождении и интерпретации понятий.)

Например, в качестве рассматриваемых объектов можно взять совокупность рабочих какой-то отрасли промышленности. Среди всех их свойств выделим только одно: эмоциональное состояние, которое можно назвать удовлетворенностью работой. В качестве моделируемых отношений выберем отношения равенства и порядка ("больше") рабочих по их удовлетворенности: считаем, что какие-то два рабочих "вступают" в отношение равенства, если их удовлетворенности в некотором содержательном плане равны, и "вступают" в отношение порядка, если, скажем уровень положительных эмоций по отношению к работе у первого рабочего больше аналогичного уровня второго.

Задачей измерения чаще всего является приписывание нашим респондентам таких чисел (подчеркнем, что результатами измерения могут быть и не числа), в которых соответствующим

образом отразились бы описанные отношения: если оказалось, что двум респондентам в результате измерения оказались приписанными одинаковые числа, то мы должны быть уверены, что соответствующие эмоциональные состояния этих респондентов одинаковы; если же первому респонденту оказалось приписанным большее число, чем второму, то у нас должна быть уверенность в том, что удовлетворенность первого респондента больше удовлетворенности второго. Ясно, что это сделать не просто – в частности, потому, что не просто оценить упомянутый выше "уровень положительных эмоций".

Аналогичные рассуждения должны быть справедливыми и для рассмотренного выше примера – для той ситуации, когда изучаемым множеством объектов служит некоторая совокупность учителей и мы рассматриваем две системы отношений между ними: отвечающие качеству их работы и материальному благосостоянию соответственно. Выбор соответствующих индикаторов по существу и означал выделение учитываемых отношений.

Желание удовлетворить рассмотренным требованиям обычно сопровождается всем тем "букетом" связанных с процессом выделения понятий и их операционализацией проблем, о которых мы упоминали в п.1.3. Но в настоящей работе нас больше волнует другой аспект того же процесса моделирования (подчеркнем, что пока речь идет о той модели, которая строится в процессе измерения) – связанный с непосредственным анализом данных.

Выделяя моделируемый при измерении фрагмент реальности и строя его модель, мы должны помнить еще об одном упомянутом там же моменте: в результаты измерения нами вкладывается еще кое-какой смысл - тот, который связан с поиском интересующей нас закономерности. Другими словами, нельзя забывать о том, ради чего осуществляется измерение, о том, какого рода закономерности нас интересуют (хотя сами закономерности мы будем находить позже, в процессе анализа данных, собранных с помощью процедуры измерения). Строя модель в процессе измерения, необходимо параллельно формировать определенные представления об изучаемом явлении – представления, адекватные последующей его формализации в процессе выбора и реализации алгоритма анализа. Естественно, при этом должно происходить абстрагирование от ряда реальных сторон этого явления. Именно это имело место, когда мы, изучая зависимость между материальным положением учителя и качеством его работы, сочли возможным использовать именно коэффициент корреляции между признаками, явившимися результатом операционализации понятий. Напомним, что это неявно вкладывалось нами в интерпретацию получаемых в результате измерения чисел. В частности, мы полагали осмысленной, содержательно интерпретируемой, структуру интервалов между числами (т.е. считали последние полученными по крайней мере по шкале интервалов). Если бы

мы предпочли, скажем, не менее известный коэффициент корреляции рангов Спирмена, то тем самым придали бы числам другой смысл – считали бы осмысленным лишь числовое отношение порядка (т.е. полагали бы, что при измерении была использована порядковая шкала).

Назовем выделенный нами фрагмент реальности *эмпирической системой (ЭС).* Таким образом, ЭС - это совокупность интересующих нас объектов вместе с системой связывающих их отношений. При этом в число таких отношений входят как те, которые мы непосредственно моделируем при измерении, так и те, которые, являясь на этапе измерения элементом интерпретации данных, будут далее использоваться в процессе анализа последних[19] . Более подробно о смысле моделируемых при построении ЭС отношений, в частности, об упомянутой интерпретации идет речь в [Интерпретация и анализ...,1987, гл.1; Толстова, 1991а, 1998].

Подчеркнем, что зачастую четкое выделение как объектов и их свойств, так и черт изучаемого явления требуют довольно высокого уровня исследовательской абстракции, и что поэтому ЭС лишь условно можно назвать фрагментом реальности. Скорее речь должна идти об определенной модели последней (той концептуально-логической модели, которая практически всегда предшествует математической). Процесс перевода всех компонент описанного фрагмента реальности на формальный, математический язык, т.е. процесс измерения, позволяет нам перейти от ЭС к некоторой *математической системе (МС).* В описанных выше ситуациях она была числовой (хотя из сказанного выше следует, что соответствующие числа совсем не обязательно являются полноценными числами в привычном всем смысле этого слова; это не имеет места, например, при использовании шкал низкого типа). Социологическим данным часто бывают адекватными и нечисловые МС (подробнее о соответствующем обобщенном понимании измерения см. [Логика социологического исследования, 1985; Толстова, 1991а, 1996в, 1998]) [20].

Заметим, что изучая интересующее нас явление, получая те или иные содержательные выводы, т.е. конкретизируя наши априорные представления о выбранной модели явления, мы пользуемся соответствующей математической теорией, т.е. свойствами задействованной МС. По существу выше, говоря о зависимости интерпретации полученных при измерении данных от того, каким методом эти данные будут анализироваться, мы говорили именно о том, что МС должна описываться интересующей нас математической теорией. Только в том случае, если последнее обстоятельство будет иметь место, можно будет применить отвечающий этой теории метод, воспользоваться разработанными в рамках этой теории положениями.

Подчеркнем, что выбирая метод анализа данных, опирающихся на какую-то математическую теорию, мы тем самым считаем эту теорию адекватной реальности. Но ответ на вопрос о том, так ли это, в социологии далеко не всегда является простым. При обосновании

соответствующей адекватности прежде всего, нужно убедиться в том, что являющиеся результатом измерения формальные объекты удовлетворяют тем свойствам, на которых базируется предполагаемая для использования математическая теория (например, аксиомам этой теории и отвечающим ей правилам вывода). После этого можно использовать известные теоремы и другие математические соотношения, выводимые в рамках упомянутой теории. Получившиеся результаты, конечно, надо будет "перевести" на содержательный язык, что отвечает шагу, в определенном смысле обратному по отношению к тому процессу формализации содержательных представлений исследователя, о котором шла речь выше[21] .

Подчеркнем, однако, что для социологических исследований подобная схема справедлива далеко не всегда. Очень часто социолог использует методы, условия применимости которых либо заведомо не выполняются, либо не проверяются. Для иллюстрации этого положения, заметим, что наиболее типичным примером свойства, которому должна удовлетворять МС при использовании многих математико-статистических алгоритмов может служить требование того, что исходные данные являются случайной выборкой из подчиняющейся определенному вероятностному закону генеральной совокупности. И такого рода свойства МС как раз очень редко проверяются (и выполняются) на практике. Тем не менее, соответствующие методы используются.

Необходимость прибегать к такого рода некорректностям объясняется, в первую очередь, тем, что математических систем, вполне адекватно отражающих те стороны реальности, которые интересуют социолога, пока придумано очень мало. Небезынтересно отметить, что в последние годы подобное положение дел привело к развитию методов изучения устойчивости разных математических алгоритмов относительно нарушений (той или иной степени) условий их применимости.

## 2.3. Основные цели применения математических методов в социологии

Использование математических методов в процессе проведения научного исследования позволяет достичь следующих целей.

Во-первых, применение математики побуждает исследователя четко сформулировать свои представления об изучаемом объекте. Правда, обусловленная сложностью социальных явлений неоднозначность соответствующей конкретизации приводит к необходимости комплексного использования нескольких методов, умелого сравнения интерпретации

соответствующих результатов и т.д. Это, с одной стороны, конечно, усложняет анализ. Но, с другой стороны, та же комплексность позволяет обогатить наши представления о реальности. Каждый подход отражает лишь какую-то одну ее грань. И только восприятие всех граней одновременно позволяет увидеть явление во всей его полноте.

Так, желая сравнить величину связи между какими-либо признаками для разных совокупностей респондентов, мы, пытаясь построить математический критерий такой связи, волей-неволей вынуждены конкретизировать свои представления о ней. Оказывается, это возможно сделать многими способами (как мы уже упоминали, только коэффициентов парной связи между номинальными признаками известно более сотни). Каждый из этих способов отражает какую-то одну сторону "истинной" связи. И лишь имея перед собой множество таких коэффициентов, мы можем понять, что есть наша связь в реальности.

Необходимость уточнения наших представлений об изучаемом явлении, вызванная потребностью использования математики, дисциплинирует исследователя и часто дает возможность ему самому лучше разобраться в том, что он изучает. Так, скажем, используя многие алгоритмы классификации для построения содержательной типологии объектов, мы вынуждены очень тщательно проанализировать наши априорные представления об искомых типах, благодаря необходимости выбрать конкретную формальную меру близости между классифицируемыми объектами (об этом см., например, [Типология и классификация в социологических исследованиях, 1982. Гл. 7]).

Во-вторых, использование математических методов позволяет четко выдержать обсужденное выше (п.2.2) абстрагирование от неисчислимого количества реальных свойств изучаемых объектов, не дает уйти в сторону от принятого исследователем понимания изучаемого явления. Конечно, в этом обстоятельстве тоже можно усмотреть и негативный аспект: любой формализм, как бы хорош он ни был, обедняет действительность; и вполне возможно, что, абстрагировавшись от чего-то, мы придем к неверным выводам из-за того, что то, от чего мы отвлекаемся, чего не принимаем в расчет, на самом деле является самым главным моментом, определяющим наше явление. Но подобных нелепостей можно избежать, если творчески, умело применять математику. Квалифицированное использование математического аппарата позволяет превратить рассматриваемую возможность последовательного абстрагирования от реальности в действенное средство помощи социологу. Ведь без использования формализма человек, к сожалению, слишком часто сбивается с единой логики рассуждения, непроизвольно подменяет одно понимание изучаемого явления другим и, естественно, в результате приходит к неверным выводам, сам того не замечая[22] .

В-третьих, с помощью математики можно получить содержательные выводы, не лежащие "на поверхности", за счет расширения круга используемых логических умозаключений. Математика по существу и предоставляет социологу возможность пользоваться всеми теми интеллектуальными достижениями, которые накопило человечество при изучении именно таких-то объектов (т.е. объектов, удовлетворяющих рассматриваемым формальным свойствам; объектов - элементов МС) и именно при таком-то понимании интересующего нас явления (т.е. при адекватности заложенной в методе модели характеру этого явления).

Так, вряд ли при изучении связи между признаками без помощи математической статистики мы сможем четко сформулировать, что такое "иметь уверенность" в неслучайности отклонения наблюдаемых частот от тех, которые должны были бы иметь место при независимости. В случае использования популярного в социологии теста "Хи-квадрат" такая уверенность появляется, когда различие между эмпирическими и теоретическими частотами достаточно большое. Что же здесь означает слово "достаточно"? Где границы большого и малого? В математической статистике ответ на такие вопросы давно получен. И формулируется он на теоретико-вероятностном языке, что вполне адекватно обычным рассуждениям социолога (более подробно соответствующая логика разъясняется в п. 2.3.1 II части настоящей книги; см. также [Толстова Ю. Н., 1990]).

Желание обойтись без математики в подобных ситуациях, вероятно, приведет нас к "изобретению" чего-то на нее похожего. А зачем изобретать велосипед? Тем более, что вряд ли у нас получится что-то лучше того, что уже придумано.

Приведем еще один пример, на наш взгляд, очень важный для социолога. Типичной задачей, решаемой исследователем в процессе анализа анкетных массивов, является задача нахождения таких сочетаний значений рассматриваемых признаков (что, очевидно, можно ассоциировать с соответствующей этим

70

сочетаниям группировкой респондентов), которые детерминируют некоторое "поведение" респондента. Скажем, "поведением" может служить голосование или неголосование за некоторого политического лидера. Результатом решения подобной задачи может служить, например, вывод о том, что среди мужчин старше 40 лет с высшим экономическим образованием и живущих в сельской местности 95 % проголосовало за рассматриваемого лидера, т.е. что респонденты с названными свойствами обладают анализируемым "поведением". Процесс решения такого рода задач обычно является чисто интуитивным. Никакой гарантии обнаружения всех требующихся групп респондентов при этом

не имеется. Более того, обычно нет гарантии и того, что мы найдем хотя бы одну группу, даже если такие группы в изучаемой совокупности имеются.

Каков же выход из подобного положения? Нам не хотелось бы все свести к необходимости привлечения на помощь ЭВМ для организации того, чего человек не может сделать просто в силу огромности требующейся работы, т.е. для простого перебора возможных сочетаний значений рассматриваемых признаков с целью выделения всех тех групп респондентов, которые обладают изучаемым "поведением" (хотя такого рода чисто механическая помощь ЭВМ, конечно, важна, к обсуждению этого обстоятельства мы еще вернемся). Такое применение ЭВМ не подразумевает использование каких бы то ни было нетривиальных логических умозаключений. Здесь же требуется несколько иной поворот дела. Математика нужна нам по существу. Дело в том, что осуществление требующегося перебора в практических ситуациях обычно бывает не под силу даже современным ЭВМ. Вот тут-то и приходят на помощь математические методы поиска требующихся сочетаний, методы, дающие определенные гарантии того, что мы такие сочетания найдем, коли они имеются в нашей совокупности. Подобные алгоритмы существуют. Некоторые из них будут рассмотрены во второй части книги – п.2.5. (например, алгоритмы типа AID) Социолог же о существовании этих методов, как правило, просто не знает. Последствия этого описаны выше.

О том, что в социологических исследованиях может использоваться разная логика рассуждений, см., например [Толстова, 1996б].

В-четвертых, не лежащие на поверхности выводы могут быть получены за счет создания возможности анализа огромных массивов информации (с которыми обычно и имеет дело социолог), учета огромного количества факторов (определяющих практически любое общественное явление). Этот аргумент "в защиту" математики обычно бывает наиболее понятным. Но указанную возможность создает не столько использование собственно математических методов, сколько применение ЭВМ (которое, однако, невозможно без применения математических алгоритмов), что само по себе для нас менее интересно: речь идет о чисто "количественной" помощи социологу, просто о более быстром проведении каких-то операций. А говоря о математическом анализе данных, нам хотелось бы в первую очередь затронуть "качественную" сторону исследовательского процесса: нас интересует, какую модель реальности мы используем, в какой степени она отражает наши представления о ней и т.д.

О роли математики в социологии говорят многие авторы (в работе [Толстова, 1991а, с. 19-20] приводится библиография). Здесь отметим очень удачную по своему жанру и исполнению книгу [Максименко, Паниотто, 1988].

В заключение настоящего раздела отметим, что без применения математического аппарата трудно обойтись при решении практически любой социологической задачи. А поскольку главной целью анализа данных является выявление статистических закономерностей, то из всех ветвей математики для социолога естественным образом на первое место выходит та ветвь, которая направлена именно на поиск таких закономерностей – математическая статистика (и, конечно, лежащая в ее основе теория вероятностей). Для того, чтобы эффективно пользоваться этой ветвью математики, необходимо понимать, что лежащие в основе математической статистики положения отражают нечто важное для социолога, и давать себе отчет в том, как, в каком виде соответствующее отражение осуществляется. Об этом и пойдет речь ниже.

# 3. АКТУАЛЬНОСТЬ ДЛЯ СОЦИОЛОГИИ ЗАДАЧ, РЕШАЕМЫХ МАТЕМАТИЧЕСКОЙ СТАТИСТИКОЙ

## 3.1. Основные задачи математической статистики с позиции потребностей социологии

Итак, главной задачей анализа данных является сжатие собранной эмпирической информации, направленное на "вычленение" скрытых в ней статистических (т.е. имеющих место "в среднем") закономерностей. Примерно так же формулируется и основная задача математической статистики. Ее методы направлены на изучение именно статистических закономерностей. Разработанные в рамках этой науки приемы позволяют выявлять "средние" тенденции, "заложенные " в исходных данных. Именно это, в первую очередь, обусловливает необходимость обращения социолога к математической статистике. Но имеются и другие причины.

Вспомним еще об одной очень остро стоящей практически перед любым исследователем-социологом проблеме – проблеме соотнесения выборки и генеральной совокупности, проблеме построения репрезентативной выборочной совокупности. Будем считать, что она в общих чертах знакома читателю[23] .

Вряд ли можно подвергнуть сомнению то, что при изучении статистических закономерностей социолога практически всегда интересует задача перенесения полученных им результатов с той совокупности объектов, которая непосредственно была обследована (с

выборки) на более широкую совокупность (генеральную). Это делает использование математической статистики еще более привлекательным для социолога: ведь с помощью соответствующих подходов можно осуществлять анализ выборочных данных именно с намерением обобщения получаемых результатов на соответствующую генеральную совокупность.

Таким образом, основные задачи математической статистики вполне адекватны задачам, которые ставит перед собой социолог. Естественно, что при решении социологических задач мы должны активно использовать все полезные для нас достижения современной науки, в том числе и математической статистики. Однако, как мы отмечали выше, при использовании соответствующих подходов в социологии и других науках, опирающихся на изучение эмпирических данных, возникают серьезные проблемы. И для того, чтобы разобраться в том, что из области математической статистики мы можем, а что не можем использовать, надо более четко понять, с какими объектами она имеет дело, и в соответствующем ракурсе более детально проанализировать, какие задачи она решает. Перейдем к более подробному обсуждению того, какие задачи позволяет решать математическая статистика и какое отношение эти задачи могут иметь к потребностям социолога.

## 3.2. Случайные величины и распределения вероятностей как основные объекты изучения математической статистики и эмпирической социологии

Основными объектами изучения для математической статистики являются т. н. случайные величины (пока – одномерные). Это функции, определенные на некоторых случайных событиях ("случайное событие" – основное понятие теории вероятностей; как известно, сам термин "вероятность" осмыслен лишь применительно к некоторому случайному событию) и принимающие числовые значения. В качестве типичного для социолога случайного события является выбор того или иного респондента. Случайными величинами могут служить признаки, определенные для этих респондентов.

Скажем, возьмем такой признак, как возраст. "Переходя" от события к событию. т.е. от одного респондента к другому (скажем, перебирая анкеты), мы будем фиксировать разные значения возраста (18, 36, 24, . .. лет), т.е. разные значения нашей случайной величины.

Случайная величина может быть многомерной – например, когда ей отвечает несколько признаков, а ее значениями являются не отдельные числа, а сочетания чисел – значений

рассматриваемых признаков. Скажем, если наряду с возрастом мы будем учитывать пол (0 - мужчина, 1 - женщина) и зарплату (в рублях), то в качестве значений нашей трехмерной случайной величины могут выступать, например, тройки чисел: (18, 0, 524), (36, 1, 1200) и т.д.

Сказанным не ограничивается определение случайной величины. Мы не упомянули самого главного – для каждой совокупности значений случайной величины должна быть определена вероятность того, что, обследуя респондентов, социолог встретит значение из этой совокупности.

Напомним, что вероятностью события называют некоторую числовую характеристику степени возможности его появления в определенных, могущих повторяться неограниченное число раз, условиях. Выше в качестве события указывался выбор респондента. О вероятности этого события говорить не будем (поскольку такая вероятность связана с правилами построения выборки, которые мы не затрагиваем). В интересующем нас случае тот факт, что случайная величина приобретает некоторое значение, сам рассматривается как случайное событие. И именно задание соответствующих вероятностей сопрягается с определением случайной величины. Условия же реализации нашего случайного события – это условия, определяющие отбор респондента.

Совокупность вероятностей встречаемости значений рассматриваемой случайной величины называется отвечающим ей распределением вероятностей, или просто ее распределением. Функция, задающая для определенных наборов значений случайной величины отвечающую им вероятность, называется функцией распределения этой случайной величины. Задать случайную величину, по существу, и означает задать соответствующее вероятностное распределение.

На практике часто используется т.н. функция плотности вероятности, определяющая, грубо говоря, вероятность встречаемости каждого значения случайной величины[24]. В качестве примера можно привести многим хорошо знакомое, часто использующееся в математической статистике нормальное распределение (которое тоже, как известно, может быть одномерным и многомерным), имеющее вид "колокола".

Подчеркнем, что самое вероятность исследователь никогда не наблюдает, в принципе не может измерить. Это – продукт нашего мышления, абстракция, идеальный конструкт[25] . Вероятность присуща генеральной совокупности, понятие которой само является абстракцией[26]. Вместо вероятности исследователь обычно имеет дело с ее выборочной оценкой – относительной частотой встречаемости соответствующего события. Косвенное обоснование целесообразности такой подмены можно усмотреть в том, что одно из известных определений

вероятности, носящее название частотного, как раз и состоит в отождествлении ее с тем пределом, к которому стремятся частоты встречаемости интересующего нас события при многократном повторении выборочных расчетов (для все новых и новых выборок).

Чтобы было возможно использование аппарата математической статистики, необходимо частотные выборочные распределения расценивать как выборочные представления генеральных распределений вероятностей. Каждое такое распределение ассоциируется со случайной величиной.

Так, например, для выборки из 10 респондентов, сведения о которой фигурируют в таблице 1, выборочное частотное распределение, отвечающее случайной величине "Удовлетворенность трудом", будет иметь вид, представленный в таблице 2.

С помощью тех же данных можно рассчитать и двумерные распределения, одно из которых приведено в таблице 3. Это - выборочное представление двумерной случайной величины, отвечающей паре признаков ("пол", "удовлетворенность трудом").

*Таблица 2.*

**Пример частотной таблицы, построенной на основе данных таблицы 1 и отражающей выборочное представление распределения случайной величины "удовлетворенность трудом".**

| Значение признака | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Частота встречаемости значения (%) | 30 | 30 | 10 | 10 | 20 |
| Выборочная оценка вероятности Р встречаемости значения | 0,3 | 0,3 | 0,1 | 0,1 | 0,2 |

*Таблица 3.*

**Пример частотной таблицы, построенной на основе данных таблицы 1 и отражающей выборочное представление распределения двумерной случайной величины ("пол", "удовлетворенность трудом").**

| Пол | Удовлетворенность | | | | | Итого |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| 1 | 3 | 1 | 0 | 1 | 1 | 6 |
| 2 | 0 | 2 | 1 | 0 | 1 | 4 |
| Итого | 3 | 3 | 1 | 1 | 2 | 10 |

В разделе 2 второй части понятие частотных таблиц будет обсуждено более подробно.

Математическая статистика позволяет находить широкий круг статистических закономерностей. Любая из них является некоторым набором параметров вероятностных распределений рассматриваемых случайных величин (одномерных и многомерных). Такого рода характеристиками являются, к примеру, разные меры средней тенденции, разброса значений случайных величин, связи между признаками и т.д. Результат, скажем, регрессионного анализа можно рассматривать как совокупность коэффициентов регрессии, которые в конечном итоге тоже являются некоторыми параметрами исходного многомерного распределения (характеристиками многомерной случайной величины) и т.д. Однако сами параметры, в той же мере, как и те вероятности, на базе которых они рассчитываются, остаются неизвестными исследователю. Вместо истинных значений параметров мы имеем только их выборочные оценки, рассчитанные на основе частотных распределений. Эти оценки называются статистиками[27].

Итак, поскольку исследователь изначально имеет дело лишь с частотами, а не с соответствующими вероятностями, то фактически исходные случайные величины предстают перед ним в весьма приближенном виде. То, что на основе выборочных данных мы рассчитываем не сами параметры распределений, а лишь их выборочные оценки (отвечающие им статистики), усугубляет степень приблизительности искомых закономерностей. Другими словами, вид закономерности, найденной для выборки, вообще говоря, будет отличаться от вида ее для генеральной совокупности. Естественно, важную роль должна играть оценка подобного различия, поскольку нас, вообще говоря, интересуют закономерности, свойственные генеральной совокупности, хотя на практике мы и имеем дело лишь с выборкой. Именно такую оценку мы и сможем сделать, пользуясь положениями математической статистики.

Основные методы, лежащие в русле математической статистики, обычно делят на две большие группы, определяемые характером рассматриваемых закономерностей и технологией их поиска: методы статистической оценки параметров (способы расчета выборочных значений параметров и перехода от выборочных значений к генеральным; математическая статистика говорит о том, каким качествам эти оценки должны обладать, чтобы как можно более походить на их генеральные прообразы, и каким образом надо строить "хорошие" статистики, отражающие известные параметры вероятностных распределений)[28] и методы проверки статистических гипотез (оценка степени правдоподобности гипотезы о наличии некоторых соотношений между случайными величинами в генеральной совокупности на основании расчета

определенных характеристик соответствующих выборочных распределений). Подробнее о сути этих задач можно прочесть, например, в [Гласс, Стэнли,1976; Паниотто, Максименко, 1982; Статистические методы анализа информации в социологических исследованиях, 1979, гл. 6][29]. Здесь подчеркнем только, что правила переноса результатов с выборки на генеральную совокупность базируются на рассмотрении некоторых выборочных статистик как случайных величин и изучении определенных параметров их вероятностных распределений (скажем, если статистика – среднее арифметическое значение какого-либо признака, то упомянутое распределение для нее получится, если представить себе бесконечное количество выборок одного и того же размера и расчет для каждой выборки этого среднего; заметим, что, как известно, дисперсия такого распределения средних обычно называется средней ошибкой выборки и очень часто используется в эмпирических исследованиях).

В решении описанных двух задач по существу и заключается поиск статистических закономерностей. Ясно, что эти задачи весьма актуальны и для социолога. Другими словами, для него является естественным такое же понимание сути искомых соотношений между наблюдаемыми величинами, какое "заложено" в математической статистике. Обоснуем это более подробно.

Должны ли случайные величины интересовать социолога? Конечно. И социолог ими фактически пользуется, не употребляя, правда, соответствующего термина. В своей специфичной ситуации исследователь использует термин "признак" или "совокупность признаков". Обращение социолога к математической статистике по существу начинается со статистической трактовки значений используемых признаков. К примеру, чаще всего, социолога интересует не тот факт, что, скажем, ученик 10 класса средней школы № 5 города N Ваня Иванов намеревается поступить в институт, а более общее явление: например, то, что среди десятиклассников рассматриваемого региона, обладающих определенными социально-демографическими характеристиками (мужчин, горожан и т.д. ), велика доля людей, намеревающихся получить высшее образование. Обобщая сказанное, можно полагать, что социолога интересует распределение долей тех объектов совокупности (десятиклассников изучаемого региона), которые обладают определенными значениями рассматриваемого признака (в нашем примере речь идет о признаке "намерение респондента"), или определенными сочетаниями значений нескольких рассматриваемых признаков (например, трех: пола, местожительства, намерения).

Первым шагом сжатия информации, содержащейся в матрице объект-признак (см. таблицу 1), как правило, является получение частотных распределений разной размерности (см.

таблицы 2 и 3). Именно с изучения таких распределений обычно начинается анализ данных.

Таким образом, в качестве случайной величины перед социологом выступает признак (набор признаков), вместо вероятностей значений случайной величины исследователь имеет дело с относительной частотой встречаемости значений признака, вместо вероятностного распределения – с частотным, вместо параметров распределения – с отвечающими им статистиками. Рассчитав интересующие его статистики, он стремится обобщить результаты на генеральную совокупность.

Итак, основной объект, изучаемый математической статистикой, – случайная величина – является основным объектом изучения и для эмпирической социологии. Основные задачи, решаемые математической статистикой служат таковыми и для социолога, занимающегося изучением собранных эмпирических данных. Но, к сожалению, оказывается, что непосредственное применение математической статистики в социологии, зачастую, бывает весьма проблематично. Условия, предполагаемые строгими теоремами математической статистики, отнюдь не всегда выполняются на практике. И тогда вместо строгой математической статистики на сцену выступает не совсем строгое ее "приближение" - анализ данных.

Поясним, что именно мешает применению методов математической статистики в социологических исследованиях. Проведем линию размежевания между математической статистикой и теми лежащими вне ее методами, которые, давая социологу возможность поиска статистических закономерностей, в то же время позволяют преодолеть соответствующие трудности.

Сразу подчеркнем, что эти трудности можно разделить на две большие группы.

Трудности первой группы типичны не только для социологии, но и для многих других наук, имеющих дело с эмпирическими данными и направленных на выявление статистических закономерностей (в числе таких наук могут быть названы биология, геология, медицина, история, психология). Именно потребности таких наук послужили толчком к развитию методов анализа данных как некой замены математико-статистических подходов для тех ситуаций, когда последние оказываются неприменимыми.

Трудности второй группы специфичны именно для социологии. Говоря о них, мы будем иметь в виду не анализ данных вообще, а анализ социологических данных. В следующем параграфе коснемся трудностей первой группы. Социологическая специфика будет затронута в разделе 5.

# 4. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА И АНАЛИЗ ДАННЫХ: ЛИНИЯ РАЗМЕЖЕВАНИЯ

Ниже, параллельно анализу рассматриваемых проблем, будем четко выделять причины, приведшие к необходимости введения наряду с термином “математическая статистика” термина “анализ данных”.

## 4.1. Проблема соотношения выборки и генеральной совокупности.

Проблемы применения математической статистики для решения интересующих нас задач начинаются с обоснования возможности использования выборочных частот в качестве хороших оценок генеральных вероятностей. Некоторые обстоятельства заставляют настороженно относиться к этому. Рассмотрим эти обстоятельства подробнее.

1) На практике нередко нарушаются условия вероятностного порождения данных.

Выше (п.3.2) мы говорили о том, что вероятность какого-либо события – это некая числовая характеристика степени возможности его появления в определенных, могущих повторяться неограниченное число раз, условиях. Определили мы и круг рассматриваемых событий – они состоят в том, что те или иные признаки принимают те или иные значения. Добавим, что понятие вероятности имеет смысл, если рассматривается "круг явлений, когда при многократном осуществлении комплекса условий S доля той части случаев, когда событие A происходит, лишь изредка уклоняется сколько-нибудь значительно от некоторой средней цифры, которая, таким образом, может служить характерным показателем массовой операции (многократного повторения комплекса S) по отношению к событию A. Для указанных явлений возможно не только констатирование случайности события A, но и количественная оценка возможности его появления. Эта оценка выражается предложением вида: … вероятность того, что при осуществлении комплекса условий S произойдет событие A, равна p” [Гнеденко, 1965. С. 15].

В социологии само определение вероятности в некоторых ситуациях может стать бессмысленным в силу ряда причин. Как правило, бывает неясно, каков тот комплекс условий, повторение которого требуется для соответствующего осмысления. Даже если некоторое смутное понимание сути этого комплекса условий у исследователя имеется, чаще всего

отсутствует уверенность в том, что этот комплекс в принципе может быть повторен и что даже при допущении его повторения мы будем иметь постоянную долю случаев реализации нашего события. В таких случаях теряет смысл гипотеза о вероятностном порождении исходных данных, принятие которой является необходимым условием корректности использования методов математической статистики[30].

Для оправдания вероятностного подхода к пониманию социологических закономерностей заметим, что упомянутая "повторяемость" предстает перед социологом в виде появления сходных ситуаций, разнесенных либо в пространстве, либо во времени. Такая точка зрения, как известно, использовалась, например, Контом, выдвинувшим в качестве основных исследовательских методов для социологии т.н. экспериментальный и сравнительный методы (в первом под экспериментом понимается исследование изменений в состоянии общества, возникающих под воздействием тех или иных потрясений, во втором – имеется в виду сравнение жизни людей, живущих в разных частях земного шара) [Конт, 1996].

Однако если в какой-то ситуации некое событие произошло, а в другой – нет, то мы практически никогда не узнаем ответа на вопрос: является ли это проявлением того, что вероятность этого события меньше единицы (реализовав много ситуаций и подсчитав долю тех, в которых наше событие свершилось, мы тем самым получим оценку соответствующей вероятности), либо же следствием того, что разные ситуации отвечают разным комплексам условий, задающих вероятность, и что поэтому вероятности нашего события в этих ситуациях различны.

Подобные рассуждения справедливы отнюдь не только для социологии. Логика развития многих наук, имеющих дело со статистическими данными, привела к необходимости "узаконивания" методов, либо не опирающихся на допущения о вероятностной природе исходных данных и, как следствие, не дающих возможности переносить результаты с выборки на генеральную совокупность), либо предполагающих подобную модель, но такую, адекватность которой невозможно проверить.

Для того чтобы как-то отделить использование математико-статистических методов в описанных ситуациях (являющееся некорректным) от их классического воплощения, для обозначения интересующих нас псевдостатистических подходов и был предложен термин "анализ данных". Это – **первая причина** появления этого термина.

2) Отнюдь не всегда бывает ясно, какова изучаемая генеральная совокупность.

Социолог имеет в своем распоряжении всего одну выборку, при том такую, принципы соотнесения которой с генеральной совокупностью часто бывают неясными. Более того,

социолог далеко не всегда уверен в том, что исследуемое им множество объектов вообще является выборкой из какой бы то ни было генеральной совокупности. Вообще, вопрос о том, что есть генеральная совокупность, по отношению к которой изучаемые объекты составляют выборку, в социологии является зачастую весьма непростым. Позволим здесь привести цитату из работы [Божков, 1988, с. 135-136], где говорится, что теоретическое обоснование и выявление качественного состава генеральной совокупности является "отнюдь не формальным и не тривиальным. ... Даже в рамках одного исследования бывают альтернативные (и множественные) решения этого вопроса. Более того, проблема определения генеральной совокупности может стать задачей или даже целью исследования. Иначе говоря, это проблема методологического, а вовсе не методико-математического характера." Мы полностью разделяем это мнение. Методы поиска закономерностей "в среднем" в подобной ситуации нельзя отнести к области математической статистики, даже если внешне они схожи с известными математико-статистическими алгоритмами. Использование этих методов в указанной ситуации было отнесено к области анализа данных. Это – **вторая причина** появления этого термина.

Таким образом, указанные сложности в применении методов математической статистики для нужд социолога в каком-то смысле преодолимы. Ниже будем полагать, что некая гипотетическая генеральная совокупность существует (хотя мы, может быть, и не знаем, какова она), и что имеющиеся в нашем распоряжении выборочные частоты – это хорошие оценки соответствующих генеральных вероятностей. Другими словами, будем считать, что вычисленное для выборки частотное распределение хорошо отражает отвечающую нашему признаку (группе признаков) случайную величину, сочтем возможным работать с этим распределением так, как правила математической статистики предписывают работать с распределением вероятностей.

Однако использование классических математико-статистических приемов соответствующего рода, зачастую, оказывается невозможным еще по нескольким причинам, также часто возникающим не только в социологии, но и в других науках, опирающихся на анализ эмпирических данных.

3) Для многих методов отсутствуют разработанные способы перенесения результатов их применения с выборки на генеральную совокупность.

Методы переноса результатов с выборки на генеральную совокупность обычно базируются на довольно серьезных теоретических результатах. Соответствующая теория не разработана для очень многих методов, интересующих социолога (например, для многих методов классификации). В результате научная ценность получаемых с их помощью выводов

оказывается весьма сомнительной: их нельзя распространить ни на какую совокупность, кроме той, для которой они были получены. Из такого положения имеется два выхода.

Во-первых, можно положиться на интуицию исследователя и считать, что результаты справедливы для некой интуитивным образом понимаемой генеральной совокупности. Так чаще всего и поступают.

Во-вторых, приложив определенные усилия, связанные с активным использованием ЭВМ, требующиеся оценки можно получить эмпирическим путем. Дело в том, что, как мы уже упоминали, правила интересующего нас переноса опираются на изучение распределений определенных статистик. Эти распределения можно искусственно создавать, рассчитывать требующиеся статистики и эмпирическим путем изучать их распределения. Другими словами, с помощью такого подхода математическая статистика из теоретической науки превращается в экспериментальную. Такой подход активно развивается на Западе, где получил название Bootstrap [Ермаков, Михайлов, 1982; Эфрон, 1988]. В последние годы он довольно часто используется и в отечественной науке.

Приведем цитату из работы [Ростовцев и др.,1997, с. 174-175]: "Классические методы статистики развивались, когда вычислительная техника еще не имела достаточного быстродействия, поэтому исследуемые статистики подбирались так, чтобы была возможность оценить их распределения. ... Современные средства анализа позволяют существенно расширить множество статистик и упростить расчеты. В частности, для оценки значимости нередко нет необходимости проводить сложные теоретические исследования распределений статистик, достаточно иметь мощный компьютер и воспользоваться методом Монте-Карло либо провести прямые вычисления вероятностей [Ермаков, Михайлов, 1982]".

4) Перенос результатов с выборки на генеральную совокупность может быть затруднен из-за осуществления "ремонта" выборки (например, ее перевзвешивания), что нередко делает социолог. Тут тоже может помочь моделирование случайных данных на ЭВМ.

Методы, для которых отсутствует строгий механизм переноса результатов с выборки на генеральную совокупность, тоже были отнесены к области анализа данных. Это – **третья причина** возникновения этого термина.

### 4.2. Отсутствие строгих обоснований возможности применения конкретных методов математической статистики. Эвристичность многих алгоритмов анализа данных

Как поиск соотношений между параметрами найденных выборочных частотных распределений, формирование соответствующих статистических гипотез и т.д., так и перенос выявленных положений на генеральную совокупность в социологии нередко затрудняется тем, что упомянутые соотношения становятся бессмысленными из-за невыполнения условий, отвечающих классическим математико-статистическим критериям. Примером может служить известное требование нормальности условных распределений при построении уравнения регрессии (напомним, что имеются в виду распределения зависимого признака, получающиеся при фиксации значения независимого). Это требование часто не выполняется, а еще чаще социолог просто не проверяет его. Последнее обстоятельство, к сожалению, нередко имеет место на практике из-за сложности проверки тех или иных условий, отсутствия соответствующего программного обеспечения, не достаточной математической грамотности социолога и т.д.

Для некоторых методов, показавших свою эффективность при решении практических задач, отсутствуют строгие доказательства корректности их использования. Это можно сказать, например, относительно применения метода регрессионного анализа к данным, полученным в результате дихотомизации номинальных признаков (об отсутствии доказательств корректности этого подхода говорят сами его авторы [Kerlinger, Pedhazur, 1973]). То же можно сказать об упомянутых нами в п. 2.3 алгоритмах типа AID – не доказано, что эти алгоритмы обязательно приведут к наилучшим “скрывающимся” в исходных данных группировкам.

Но, несмотря на все сказанное, как-то анализировать, изучать данные нам нужно. И... методы используются, несмотря на их некорректность. Это делается и в социологии, и во многих других науках, так или иначе ориентированных на получение теоретических выводов на базе наблюдения большого количества данных (биологии, психологии, геологии, медицине и т.д.). Потребности практики обусловили необходимость обращения исследователей к таким методам, жизнь заставила их мириться с соответствующими некорректностями. Более того, в математике начали вырабатываться своеобразные подходы, направленные не на разработку методов, корректных в той или иной сложной реальной ситуации, а на анализ того, в какой мере могут быть нарушены условия применимости известных методов, чтобы результаты их применения "не слишком" искажали реальность.

"Классические" математические статистики поначалу в принципе отвергали такой подход. Но жизнь взяла свое. И для обозначения совокупности таких некорректных методов, для отделения их от строгих математико-статистических подходов, был введен термин "анализ

данных". Итак, мы рассмотрели **четвертую причину** введения основного интересующего нас термина.

Отметим, что из-за невозможности использования апробированных схем математической статистики для такого рода методов, особое значение для них приобретает проблема обоснованности получаемых с их помощью выводов. От традиционных математико-статистических критериев качества здесь зачастую переходят к требованиям экстремальности некоторых специальным образом построенных критериев-функционалов. Здесь особенно остро стоит вопрос о выделении "точек соприкосновения" содержания задачи и математического формализма, чему в разделе 5 мы уделим большее внимание. Соответствующие положения послужат основой для выделения тех специфических черт, которые отличают анализ социологических данных от анализа данных вообще.

Перейдем к рассмотрению других моментов, мешающих использовать многие математико-статистические построения как в социологии, так и в других науках, опирающихся на анализ статистических эмпирических данных.

## 4.3. Использование шкал низких типов

Проблемы с использованием в социологии традиционных математико-статистических методов возникают также в связи с тем, что интересующие социолога данные, как правило, бывают получены по шкалам низких типов. Определения понятий "тип шкалы", "шкала низкого (соответственно, высокого) типа" мы заимствуем из теории измерений (ее положения описаны, например, в [Суппес, Зинес, 1967]; более простое, рассчитанное на социологов, изложение основных ее принципов можно найти в [Толстова, 1990 а, б; 1998]. Попытаемся понять, что такое шкала низкого типа хотя бы на интуитивном уровне[31].

К шкалам низкого типа обычно относят шкалы, позволяющие получать "числа", очень не похожие на те действительные числа, к которым мы привыкли, осваивая курс школьной математики. Эта непохожесть означает невозможность работать с этими числами по обычным правилам арифметики. К шкалам же высокого типа причисляют те, с помощью которых получаются числа, в достаточной мере похожие на действительные числа, т.е. такие, с которыми позволено делать почти все, что мы привыкли делать с числами. Шкалами низкого типа обычно считают шкалы, называемые в литературе номинальными и порядковыми, а шкалами высокого типа – интервальные и шкалы отношений (в теории измерений известны и другие шкалы как

низкого, так и высокого типов). Шкалы низкого типа (и получаемые с их помощью данные) часто называют также качественными, а шкалы высокого типа (и соответствующие данные) – количественными, или числовыми.

Мы отрицательно относимся к введенным в предыдущем абзаце терминам "низкий", "высокий" и особенно – "качественный" и "количественный". И не потому, что любим терминологические споры, а потому, что, на наш взгляд, описанная терминология не может не увести использующего ее социолога в сторону от правильного (с нашей точки зрения и с точки зрения исследователей, работающих в рамках теории измерений) понимания шкалы и, как следствие, понимания того, что можно делать с полученными с ее помощью шкальными значениями, как можно интерпретировать результаты анализа таких данных. О соответствующих соображениях см. [Толстова, 1990 а, б; 1998]. Тем не менее, будем пользоваться описанной выше, принятой для социологической литературы терминологией, стараясь, однако, приблизить описание номинальных и порядковых шкал к тем представлениям о них, которые кажутся нам правильными (мы имеем в виду достаточно тщательное отслеживание того, какую реальность мы отражаем в числах при использовании той или иной шкалы).

Итак, *номинальной* шкалой мы называем такую шкалу, с помощью которой стремимся отразить в числах только некоторое отношение равенства-неравенства между изучаемыми объектами. Типичным признаком, значения которого обычно получаются именно по номинальной шкале, является профессия респондента. Если одному объекту (респонденту) приписано значение "3" (отвечающее, скажем, профессии "токарь"), а другому – значение "4" (отвечающее профессии "пекарь"), то, имея в руках эти числа, мы можем быть уверенными в том, что рассматриваемые объекты в интересующем нас отношении различны (респонденты имеют разные профессии), но больше ничего мы о них сказать не можем. Говоря точнее, мы не можем использовать какие-то другие свойства чисел для формирования содержательных выводов : мы не знаем, больше ли один из респондентов, чем другой, или меньше (как 4 больше 3); можно ли говорить о том, что различие между какими-то двумя объектами равно различию между некоторыми двумя другими объектами (как различие между 4 и 3 равно различию между 3 и 2) и т.д. Другими словами, интерпретируя так или иначе полученные шкальные значения, мы можем пользоваться только теми свойствами чисел, за которыми "стоят" содержательные свойства изучаемых объектов (из теории измерений следует, что это положение не всегда имеет смысл считать верным, но здесь мы не можем остановиться на этом более подробно). В случае номинальной шкалы содержательные свойства "стоят" только за равенством и неравенством

чисел.

При использовании *порядковой* шкалы мы ставим своей целью отобразить не только некоторое отношение равенства-неравенства между реальными объектами, но и какое-то содержательное отношение порядка между ними. Обычно в качестве примеров признаков, значения которых можно считать полученными по порядковой шкале, приводят признаки, отвечающие заданным в анкете вопросам типа: "Удовлетворены ли Вы Вашей работой (ходом реформ, президентом РФ, качеством рыночных продуктов и т.д.)?" с традиционным веером из пяти (трех, семи и т.д.) вариантов ответов от "Совершенно не удовлетворен" до "Вполне удовлетворен", которым ставятся в соответствие числа от 1 до 5 (от 1 до 3, от 1 до 7, от -3 до +3 и т.д.). Здесь мы при осуществлении шкалирования ставим своей целью отобразить в числах не только отношение равенства респондентов по их удовлетворенности заданным исследователем объектом, но и отношение порядка между респондентами по степени "накала" их эмоций, направленных в адрес этого объекта. И если окажется, что одному респонденту приписано число "2", а другому - "4", то мы будем полагать, что упомянутый "накал" второго респондента не просто не равен "накалу" первого, но больше такового[32] (ясно, что здесь речь идет по существу о том отражении эмпирической системы в математическую, о которой мы говорили в п. 2.2).

Естественно, что для "чисел", полученных по шкалам низких типов, не будет иметь смысла большинство традиционных, привычных нам операций с числами. Точнее – будут бессмысленными практически все содержательные выводы, которые было бы естественно сделать из тех или иных числовых соотношений. Так, вряд ли найдется человек, усматривающий что-то рациональное в утверждениях типа: "среднее арифметическое значение профессий для рассматриваемой совокупности респондентов равно 3,2, и оно меньше аналогичного среднего значения для другой совокупности, равного 3,9" (надеемся, что определение среднего арифметического читателю знакомо). Ведь совершенно ясно, что упомянутые числа бессмысленны. Что значит величина 3, 2? То, что некий средний, наиболее типичный респондент на 20% является токарем, а на 80% - пекарем? Бред такого использования традиционной статистической характеристики (среднего арифметического) очевиден.

Вернемся к проблеме соотнесения принципов математической статистики с потребностями социологии.

Итак, интересующие социолога данные чаще всего бывают получены по шкалам низких типов – номинальной или порядковой. Случайные же величины, с которыми имеет дело математическая статистика, обычно предполагаются числовыми, т.е. такими, значениями

которых служат обычные действительные числа. Таким образом, с "социологическими" числами мы не имеем права поступать, как с обычными числами, с "математико-статистическими" же числами можем делать все, что угодно. Правда, здесь следует оговориться, что большая часть результатов математической статистики пригодна для применения к данным, полученным по *интервальным* шкалам. Соответствующие шкальные значения "почти" похожи на всем привычные действительные числа, но все же таковыми не являются. Они отображают в числовые отношения не только некоторые эмпирические отношения равенства и порядка, но и структуру эмпирических интервалов – отношения равенства и порядка для расстояний между объектами. Интервальные шкалы часто называют числовыми, хотя это и не совсем точно. Ниже мы не будем делать различия между шкальными значениями, отвечающими интервальной шкале, и всем привычными действительными числами.

Из-за различия в типах шкал, используемых математической статистикой и социологией, перенос того, что мы получаем в математической статистике, в социологическую практику часто оказывается невозможным. Часто, но не всегда.

Дело в том, что в математической статистике имеются и такие разделы, которые посвящены анализу частотных распределений для номинальных и порядковых признаков. Но, используя соответствующие результаты, мы тем самым не только полагаем, что выборочные частоты хорошо приближают генеральные вероятности (ср. п.4.1), но и делаем ряд других допущений, на которые опираются рассматриваемые математико-статистические утверждения.

Одним из самых главных с точки зрения важности его роли для социолога является предположение о том, что за анализируемыми номинальными и порядковыми признаками как бы "стоят" некоторые числовые переменные. Выполнение этого предположения в социологических задачах часто является весьма проблематичным. Многие же методы математической статистики опираются на это предположение (среди них самый популярный у социологов метод измерения связи между номинальными переменными, метод, основанный на критерии Хи-квадрат). Здесь мы не будем вдаваться в подробности. Для нас важно констатировать, что использование шкал низкого типа очевидным образом затрудняет применение классической математической статистики при решении социологических задач.

Подчеркнем также, что вопрос о принятии (непринятии) рассмотренного предположения самым непосредственным образом связан с нашими содержательными представлениями о том, что скрывается за понятием "признак", – например, с нашей интерпретацией восприятия респондентом предлагаемых ему вопросов. Это, конечно, имеет прямое отношение к проблеме социологического измерения, которую мы здесь, вообще говоря, не рассматриваем, но

пользуемся случаем лишний раз подчеркнуть специфичную для социологии органическую связь между измерением и анализом данных. Кроме того, обратим внимание читателя на то, что тот же вопрос тесно связан с проблемой соотнесения модели, "заложенной" в методе, с содержательным характером задачи. К этому мы еще вернемся в следующем разделе при рассмотрении соответствующих аспектов анализа социологических данных.

Имеются и другие возможности использования математической статистики для изучения данных, полученных по шкалам низких типов. Мы имеем в виду не ставшую еще общеизвестной новую ветвь этой науки, носящую название статистики объектов нечисловой природы [Орлов, 1985]. Однако наработок, осуществленных в этой области, при всей их значимости, пока не достаточно для того, чтобы удовлетворить потребности практики.

Отметим, что рассматриваемые трудности присущи процессу поиска статистических закономерностей отнюдь не только в социологии. Т. н. качественные данные встречаются и во многих других науках. Методы, позволяющие осуществлять указанный поиск, также были отнесены к понятию "анализ данных". Иными словами, необходимость анализа "чисел", полученных по шкалам низких типов, послужила **пятой причиной** "рождения" названного понятия.

Итак, говоря о необходимости специального рассмотрения "неправильных" с точки зрения математической статистики методов поиска статистических закономерностей, мы пока оправдываем такую необходимость в основном потребностями многих наук. Анализ же социологических данных обладает рядом специфических черт, которые выделяют его из анализа данных вообще. И специфичные моменты процесса поиска статистических закономерностей именно в социологии связаны, в первую очередь, с тем выделением "точек соприкосновения" содержания задачи и математического формализма, о котором мы упоминали выше. И это связано с **шестой причиной** (может быть, самой важной для социолога) рождения понятия "анализ данных", причиной, обусловленной сложностью изучаемых с помощью анализа данных явлений – необходимостью постоянного вмешательства исследователя в процесс анализа.

Рассмотрим соответствующие вопросы, касающиеся именно социологии, более подробно.

# 5. СПЕЦИФИКА ИСПОЛЬЗОВАНИЯ МЕТОДОВ АНАЛИЗА ДАННЫХ В СОЦИОЛОГИИ

## 5.1. Необходимость соотнесения модели, "заложенной" в методе, с

## содержанием задачи

Выше мы уже говорили о том, что любой математический метод предполагает адекватной реальности определенную модель того явления, которое с помощью этого метода изучается. Но любая модель – это лишь некоторое приближение к действительности. Рассмотрим более подробно вопрос о достаточности такого приближения для социологических задач анализа данных.

Одним из проявлений трудностей с формализацией наших представлений о социальных явлениях является то, о чем мы уже упоминали: если для решения какой-то задачи существует некоторый математический метод, то этот метод практически никогда не бывает единственным. Примером могут служить уже самые простые характеристики одномерных распределений. Так, вообще говоря, существует много мер средней тенденции (и разброса) частотного распределения значений любого признака. Выше уже говорилось о том, что для измерения связи даже между двумя номинальными признаками могут служить более сотни известных из литературы коэффициентов соответствующего плана. Еще большее разнообразие присуще сложным методам изучения многомерных распределений[33]. И за каждым методом "стоит" свое понимание изучаемого явления (средней тенденции, разброса, связи и т.д.).

Какой метод выбрать? Как сравнивать результаты применения разных методов? Эти и другие подобные вопросы встают практически перед каждым исследователем. И любой социолог, использующий хотя бы самые элементарные математические методы (скажем, рассчитывающий среднее арифметическое значение, моду, медиану какого-либо признака), зачастую фактически дает ответы на вопросы такого рода, даже если он об этом и не задумывается (а, к примеру, при использовании какого-либо относительно сложного метода выбирает с помощью ЭВМ вариант "по умолчанию").

Все сказанное обусловливает особую остроту для социологии вопроса об адекватности модели, заложенной в том или ином методе, содержанию решаемой с помощью этого метода задачи (точнее, концептуальным представлениям исследователя о ее сути). Реализация процесса соответствующего соотнесения – задача социолога. И здесь вряд ли помогут советы представителей других наук. Ведь решение этой задачи требует обеспечения естественности используемого математического языка; вычленения из живой реальности моделируемых с помощью математики фрагментов; четкого выделения таких элементов используемых алгоритмов, которые имеют непосредственный "выход" на содержательные представления

социолога об изучаемом явлении. Приведенное утверждение является достаточно общим и, вероятно, может показаться в какой-то степени очевидным. Однако лишь задавшись целью обязательного сопряжения формализма и содержания, можно прийти к тем многочисленным и (как нам представляется), далеко не столь тривиальным, утверждениям, которые можно считать конкретизацией высказанного положения применительно к реальным интересующим социологов методам.

Приведем несколько примеров.

Начнем, казалось бы, с самого простого – с расчета мер средней тенденции. В математике известно бесконечное количество таких мер. В руководствах, ориентированных на социолога, обычно рекомендуют три из них – те, которые были названы нами выше – среднее арифметическое, медиану, моду. Сейчас мы не будем принимать в расчет то, что, как хорошо знает каждый социолог, далеко не для всех шкал могут быть использованы две первые меры. Рассмотрим случай, когда тип шкалы нас не ограничивает в выборе среднего (предположим, например, что мы имеем дело с интервальными шкалами). Для того, чтобы показать, что такой выбор может диктовать нам содержание задачи, позволим себе описать несколько эксцентричный пример, приведенный нами в [Толстова, 1990а, с. 62-63].

Опишем некоторую задачу о моде в житейском смысле этого слова. Предположим, что модельер должен определить, какая длина должна быть у очередной модели женских юбок, выпускаемых какой-то фабрикой, и для этой цели опрашивает женщин рассматриваемого региона, просит их указать "любимую" длину. Если мы в качестве длины, рекомендуемой фабрике, укажем медиану соответствующего распределения, то тем самым окажемся перед риском выпустить неходовой товар: половина женщин решит, что юбка для них слишком коротка, а половина – что чересчур длинна. Покупать продукцию фабрики никто не захочет. А вот если в качестве меры средней тенденции мы используем моду, то удовлетворим женщин, выразивших наиболее часто встречающееся мнение.

Коротко укажем на другие известные из литературы примеры. Терстоун, предлагая свой хорошо известный (см., например, [Толстова, 1998]) метод построения шкалы для измерения установки, рекомендовал на последнем этапе процедуры, при расчете приписываемого каждому респонденту итогового балла, использовать медиану в качестве среднего значения весов тех суждений, с которыми этот респондент согласился (а не среднее арифметическое, хотя с формальной точки зрения его в данном случае можно было бы посчитать; правда, здесь мы используем определенный взгляд на тип получающихся шкал, который требует специального обсуждения).

В некоторых конкретных ситуациях может возникнуть потребность использования совершенно иных мер средней тенденции. Так, в [Дэйвисон, 1988] рассматривается задача изучения пространства восприятия респондентами некоторых объектов с помощью многомерного шкалирования. Предлагается способ построения матрицы близости между объектами на основе своеобразного опроса респондентов. И для усреднения соответствующих мнений рекомендуется использовать среднее геометрическое.

Приведем еще один пример, где речь идет о более сложном (по сравнению с расчетом средних) методе анализа данных. Предположим, что мы хотим построить типологию изучаемых объектов, используя для этого какой-либо из алгоритмов многомерной классификации (напомним, что в соответствии с этими алгоритмами каждый классифицируемый объект задается как точка некоторого признакового пространства). В таком случае выбор алгоритма должен определяться нашими априорными представлениями об искомых типах. Так, если мы считаем, что каждый тип может быть представлен неким "центральным" объектом, вокруг которого "кучкуются" другие объекты того же типа (т.е. если все однотипные объекты близки друг к другу одновременно по всем рассматриваемым признакам и, вследствие этого, центральный объект может служить как бы "олицетворением" типа), то мы должны выбрать какой-либо из алгоритмов, направленных на поиск круглых "сгущений" в рассматриваемом признаковом пространстве. Если же мы отождествляем каждый искомый тип с тем, какова форма зависимости какого-либо из рассматриваемых признаков от остальных, то подобные алгоритмы в принципе становятся неприменимыми. В таких случаях надо использовать методы, позволяющие искать "длинные" скопления точек в признаковом пространстве, "олицетворяющие" упомянутые зависимости.

Более обстоятельное описание подобных ситуаций можно найти, например, в работах [Патрушев и др., 1980; Типология и классификация в социологических исследованиях, 1982; Математические методы анализа и интерпретация . . ., гл. 1], где подробно говорится о той априорной модели, которую должен сформировать исследователь, желающий решать задачу типологии тех или иных объектов с помощью методов многомерной классификации (речь идет об априорных представлениях об искомых типах и о том, что, не имея таких представлений, исследователь рискует получить нелепые результаты, поскольку в таком случае математика не может выполнять функции "орудия труда" социолога).

Ясно, что социолог должен уделять большое внимание анализу моделей, заложенных в используемых им методах. И это – одна из причин присутствия термина "социологический" в названии нашей работы. Но существуют и другие.

## 5.2. Связь разных этапов исследования друг с другом

Для того, чтобы использование математического языка обладало той естественностью, о которой шла речь выше, необходимо, чтобы применение математики было буквально вплетено в логическую канву исследования. Математика должна служить "орудием труда" социолога, а не играть роль инструмента "пришлепывания" к исследованию модного "бантика", не очень-то вяжущегося со всем остальным (что, к сожалению, очень часто бывает на практике). Для достижения этой цели недостаточно того сопряжения формализма и содержания, о котором мы только что говорили. Чтобы не оставлять за математикой лишь роль средства придания некоторого наукообразия работе социолога, необходимо учитывать, что корректность использования математического аппарата на любом из этапов исследования тесно связана с принципами реализации других этапов, в том числе и таких, в которых не задействованы никакие математические методы. Это требование конкретизируется в виде целого ряда положений. Из-за недостатка места мы упомянем только два, сопроводив их примерами использования в социологии сравнительно сложных методов анализа данных.

**Первое** – о связи измерения и анализа его результатов. В п. 1.3 мы уже упоминали о целесообразности сопряжения самого понятия статистической закономерности не только с выбором собственно алгоритма ее нахождения, но и с тем, что такому выбору предшествует и, в первую очередь, с формированием используемых понятий и способа их операционализации. А это – стадии процесса измерения. Там же, а также в п. 2.2 речь шла о том, что выбор конкретного алгоритма анализа и интерпретация результатов измерения взаимно обусловливают друг друга. Конечно, серьезное обсуждение указанной связи невозможно без конкретизации соответствующих положений для тех или иных используемых в социологии алгоритмов, что требует рассмотрения последних и не входит в число наших задач. Тем не менее, приведем небольшой пример, чтобы пояснить, что мы имеем в виду.

Соответствующие соображения уже были описаны нами в [Типология и классификация …, 1982]. Осуществляя типологию респондентов на основе данных об их бюджетах времени, мы стоим перед выбором: можно считать, что количества минут, затраченных тем или иным респондентом на какие-то виды деятельности, могут нами восприниматься с точки зрения различий разностей между ними (например, можно считать осмысленными, естественным образом интерпретируемыми выражения типа 120-80=50-10); можно полагать, что нам важна

только структура времяпрепровождения человека (и, как следствие, учитывать не указанные разности, а то, что 120 в полтора раза больше, чем 80, а 50 – в пять раз больше, чем 10), а можно "видеть" в рассматриваемых количествах минут лишь порядок их расположения по величине (в таком случае указанные выше разности и отношения для нас становятся содержательно бессмысленными; о соответствующих числах мы можем сказать только, что 80<120, 10<50). Каждый вариант означает свою интерпретацию результатов измерения. Что именно мы выберем – зависит от нашего априорного понимания типа респондента (и, значит, от реализации еще одного этапа исследования – первичного формирования проверяемых гипотез). Но наше решение определит то, какой алгоритм классификации мы выберем для построения требующейся типологии.

Существует много других причин, обусловливающих неразрывную связь между измерением и анализом данных. В социологии практически никогда нельзя провести четкую границу между этими двумя понятиями. Так, наиболее интересные для социолога переменные чаще всего являются латентными, их значения не поддаются непосредственному наблюдению. Такие переменные измеряются не в процессе первичного сбора (наблюдения) данных, а в процессе анализа некоторой полученной в результате непосредственного наблюдения информации (для этого используются такие методы, как факторный, латентно-структурный анализ, многомерное шкалирование, методы парных сравнений, методы одномерного шкалирования Терстоуна, Лайкерта и т.д.). Напротив, многие методы анализа интересуют исследователя, в первую очередь, как результаты определенного рода измерения некоторых переменных. К примеру, именно с соответствующей точки зрения социолог часто интерпретирует результаты многомерной классификации: номер класса рассматривается им как значение переменной, которую можно было бы назвать "тип объекта".

Неразрывность двух проблем – построения т.н. признакового пространства (т.е. выявления способа описания исходных объектов) и выбора алгоритма анализа соответствующих данных – косвенно подтверждается наличием довольно большого количества работ, посвященных предложению методов одновременного решения этих проблем для некоторых классов содержательных задач [Браверман и др., 1974; Применение факторного ..., 1976; Типология и классификация ..., 1982].

Сказанным мы, к сожалению, здесь вынуждены ограничить рассмотрение проблемы связи измерения и анализа данных, хотя рассматриваемая проблема весьма важна и с теоретической, и с практической точки зрения, и требует более глубокой проработки.

**Второе** - о зависимости интерпретации результатов применения метода от

концептуальных установок исследователя, от стоящих перед ним целей. Для примера вспомним наше обсуждение возможных подходов к построению многомерной типологии изучаемых объектов с помощью разных алгоритмов классификации (п.5.1). Если мы считаем, что каждый тип может быть представлен неким "центральным" объектом, к которому примыкают другие объекты того же типа и выбираем алгоритм, направленный на поиск круглых "сгущений" в рассматриваемом признаковом пространстве, то для интерпретации результатов классификации можно будет рассчитывать координаты центра тяжести каждого из найденных классов. Этот центр, как мы упоминали, можно считать "олицетворением" класса. Если же мы отождествляем каждый искомый тип с тем, какова форма зависимости какого-либо из рассматриваемых признаков от остальных, то подобная интерпретация становится неприменимой. В таких случаях для интерпретации надо искать упомянутые зависимости.

Упомянем также пример, уже описанный нами в [Математические методы анализа …, 1989]. В этом примере в процессе рассмотрения той же задачи построения типологии респондентов рассказывается, каким образом представления социолога об искомых типах позволяют корректировать результаты формальной классификации с целью превращения ее в содержательно интерпретируемую типологию.

Будем считать, что приведенных примеров достаточно для того, чтобы сформировать хотя бы самые приблизительные представления о том, что мы имеем в виду, говоря о необходимости соотнесения всех этапов исследования друг с другом. И представляется совершенно очевидным то, что такое соотнесение может быть осуществлено только самим социологом. Ведь оно по существу означает определенную целостность, неразрывность всего социологического исследования.

## 5.3. Другие методологические принципы анализа социологических данных

Выше мы сформулировали два основных методологических принципа, соблюдение которых является необходимым для того, чтобы использование математики было эффективным: сопряжение формализма и содержания и органическая связь всех этапов исследования друг с другом. Можно было бы говорить еще о целом ряде подобных требований, носящих более частный характер: необходимость выполнения некоторых принципов измерения интересующих социолога показателей; обеспечения определенной однородности той совокупности объектов,

на которой "действует" наша предполагаемая закономерность; соблюдения некоторых принципов интерпретации результатов применения метода; выполнения определенных правил комплексного использования целой серии методов при решении практически любой социологической задачи и т.д. (некоторая "сводка" подобных принципов дана нами в [Толстова, 1991а, б]).

Раскрытие каждого из названных принципов требует серьезного рассмотрения. Все они многоаспектны, имеют сложную структуру. Их практическая реализация требует достаточно глубокого анализа концептуальных представлений социолога об изучаемом явлении, для чего требуется четкая формулировка самих этих представлений.

Так, говоря об *измерении*, мы должны давать себе отчет в том, какие именно элементы реальности собираемся отобразить в тех или иных математических конструктах (чаще всего - в числах); какова наша модель восприятия респондентом предлагаемых ему объектов (суждений и т.п. ); какая именно интерпретация этих конструктов будет нами использоваться при их анализе и т.д. [Толстова, 1998]

Обеспечивая *однородность* подвергаемой анализу совокупности данных о наших объектах, необходимо задуматься о том, имеем ли мы право для всех интересующих нас респондентов использовать один и тот же инструмент измерения и одинаковым образом интерпретировать результаты последнего; можем ли мы считать, что формальный вид искомой закономерности должен быть одним и тем же для всей выборки; можем ли мы одинаковым способом интерпретировать результаты анализа и т.д. [Толстова, 1986, 1991а].

*Интерпретируя* результаты применения того или иного алгоритма анализа мы должны обеспечивать, чтобы эта интерпретация не противоречила интерпретации исходных данных; чтобы при ее осуществлении по возможности компенсировались бы те недостатки формализма, которые волей-неволей мы вынуждены были игнорировать при измерении и выборе метода анализа ("идеальная" формализация того, что интересует социолога, как правило, бывает невозможна) и т.д. [Интерпретация и анализ … . Гл. 1; Толстова, 1991а]

Продумывая вопрос об адекватности тех или иных методов измерения и анализа данных, понимая, что все они не в полной мере отражают то, что нужно социологу, последний часто приходит (или должен приходить) к выводу о том, что достаточно полное отражение интересующей его картины реальности требует *комплексного использования разных методов*. За каждым - свои плюсы и минусы. А будучи примененными в комплексе друг с другом, они могут дать вполне адекватное представление о действительности. Но здесь встает множество

вопросов, связанных с глубоким анализом модели, заложенной в каждом методе, с разработкой принципов сравнения разных методов друг с другом и т.д. [Толстова, 1991а]

Полагаем, что сказанного достаточно для того, чтобы читателю стало ясно, почему (и в каком смысле) в заглавии нашей книги мы "привязываем" анализ данных именно к социологии.

* * *

Итак, мы в самых общих чертах описали, что такое "анализ социологических данных". При этом мы не только активно использовали то, что о соответствующих вопросах говорится в литературе, но и изложили свое видение ряда положений. Последнее в особой степени касается роли термина "социологический" в интересующем нас словосочетании.

Выше коротко раскрыта роль методов анализа данных в социологии и рассмотрены основные методологические принципы их использования при изучении общественных процессов. Конечно, все изложенное раскрывает суть анализа социологических данных действительно лишь "в самых общих чертах". Поэтому, вероятно, не все сказанное выше стало читателю полностью понятно; отдельные положения, может быть, показались очевидными либо, напротив, слишком "заумными", оторванными от реальности.

Наше убеждение состоит в том, что все приведенные соображения имеют самое непосредственное отношение к практике, к обеспечению хорошего научного уровня любого эмпирического социологического исследования. И каждое сформулированное выше утверждение становится весьма нетривиальным, когда дело доходит до его воплощения в жизнь. Но показать это, равно как и разъяснить более подробно то, что, возможно, осталось неясным читателю, можно только на реальных примерах. Необходимы: рассмотрение реальных социологических задач; демонстрация того, как их решению может способствовать математический аппарат; подробный анализ процесса сопряжения каждого метода с концептуальными представлениями исследователя и т.д. В определенной мере об этом пойдет речь во второй части (особый упор будет сделан на анализ моделей, заложенных в рассматриваемых методах).

**ПРИМЕЧАНИЯ К ЧАСТИ I.**

[1] Вероятно, здесь можно было бы говорить практически обо всех теоретических построениях, поскольку даже самые абстрактные логические рассуждения интересующего нас плана, так или иначе, прямо или опосредованно, в конечном итоге базируются на какой-то переработке сознанием автора неких фактов. Об этом красноречиво говорит творчество практически всех великих социологов, "перелопативших" огромное количество эмпирического материала: Маркса, который, по словам Энгельса, оставил после себя только по русской статистике два кубометра материалов; Дюркгейма, основной целью которого было подведение под социальную науку эмпирической базы, для которого понятие социального факта было ключевым, а множество таких фактов выступало в качестве предмета социологии; Вебера, который, изучая римскую аграрную историю, ввел в науку термин "эмпирическая социология". Вероятно, здесь целесообразно также отметить, что у древних греков даже математика была эмпирической наукой. Так, пифагорейцы впервые получали столь знакомые нам теперь результаты на базе экспериментов, числа мыслились зримо, в виде камушков и т.д. [Волошинов, 1993, с.117]. Дело дошло до того, что Платон (живший, как известно, лет через 200 после Пифагора) упрекал пифагорейцев за излишний эмпиризм [Жмудь, 1994, с. 220]. Небезынтересно отметить, что в наше время математика после двух тысячелетий пребывания в классическом дедуктивном виде снова приобретает черты экспериментальной науки, причиной чему является необходимость удовлетворить потребности таких наук, как социология (см. п.4.2 части I).

Несмотря на сказанное, все же точка зрения, в соответствии с которой все социологические утверждения базируются на анализе фактов, является спорной. Так, в [Монсон, 1992, с. 31] говорится о том, что вряд ли, ссылаясь на эмпирические факты, можно ответить на вопросы типа: "Насколько свободно и сознательно мы создаем наши социальные связи?", "Является ли общество непредсказуемым и изменчивым результатом толкований и действий отдельных людей, или это структура, которая создается и воссоздается независимо от желания и ведома отдельных ее участников?" Но, наверное, если рассматривать вопросы более частного порядка, то от необходимости анализа эмпирических данных мы все же заведомо никуда не уйдем. Мы здесь не хотели бы более глубоко обсуждать вопрос о понятии эмпирического факта, его соотношении с наблюдаемыми данными, его роли в построении социологической теории. "Уйдем" от проблемы, посчитав, что предметом нашего рассмотрения являются не любые социологические задачи, а лишь такие, которые можно отнести к т. н. эмпирической социологии

(хотя смысл этого термина в литературе тоже понимается неоднозначно).

[2] Поскольку наши данные - это лишь некоторая модель реальности, а любую модель еще надо построить, используя определенные научные представления, здесь представляется уместным провести параллель с тем, что ". . . научный факт есть определенный итог познавательного процесса, а не его начало" [Ядов, 1998].

[3] Напомним, что "цифра" – это просто значок, который, вообще говоря, может обозначать что угодно, хотя чаще всего используется для обозначения чисел, а "число" - это строго определенный математический конструкт, обладающий общеизвестными свойствами. Понятие числа /целого, положительного, мнимого и др. / в математике обычно задается аксиоматически, при этом в качестве аксиом выступают известные положения об упорядоченности чисел, о существовании для них операции сложения и т.д. Из п. 4.3 видно, что не зря мы сейчас вспоминаем эти определения; увидим, что для социолога термин "число" может скрывать за собой и несколько иной смысл и что именно для анализа социологических данных изучение этого смысла носит первоочередной характер.

[4] Подчеркнем, что под термином "объект" здесь мы имеем в виду единицу наблюдения - предприятие, респондента и др. Следует отличать такое использование этого термина от употребления его в сочетании "объект исследования", под которым понимается "все то, что явно или неявно содержит социальное противоречие и порождает проблемную ситуацию... то, на что направлен процесс познания" [Ядов, 1998]. Объектами исследования могут быть, например, отрасль народного хозяйства, коллектив какого-либо завода и т.д.

[5] Раскрытию термина "понятие" посвящено огромное количество работ. Правда, в основном они принадлежат смежным с социологией областям знаний – философии, лингвистике, психологии, психолингвистике, герменевтике и т.д. – и, вероятно, поэтому соответствующие наработки крайне редко используются социологами в процессе эмпирических исследований. Об этом остается только сожалеть.

Более того, в эмпирической социологии, как правило, не используются и "родные" результаты, полученные именно социологами. Здесь, в первую очередь, необходимо вспомнить об идеальных типах Вебера. Нельзя сказать, что соответствующие представления вообще не учитываются (например, в [Голод, 1996] идет речь об идеальных типах современной семьи). Но, на наш взгляд, использование их требуется практически в любом социологическом исследовании, что явно не имеет места. Например, при изучении, скажем, факторов, определяющих уровень успеваемости студентов, по нашему мнению, прежде, чем составлять анкету, надо сформировать представление об идеальных типах "хорошего" и "плохого"

студента. Подобные представления целесообразно использовать и на других этапах исследования: при выборе метода анализа данных, интерпретации результатов его применения и т.д.

Не будем приводить пространную библиографию, посвященную раскрытию термина "понятие". В качестве наиболее "свежих" и достаточно фундаментальных работ назовем [Войшвилло, 1989; Степанов, 1990; Кузнецов, 1997].

[6] Проблема операционализации понятий (т.е. построения их эмпирических референтов) сложна и многогранна. Мы не ставим своей целью подробное ее рассмотрение (хотя мы не можем ее совсем отбросить; и в той мере, в какой такое рассмотрение является необходимым для описания принципов грамотного использования методов анализа данных, оно осуществляется нами в п.1.3 и п.2.2). Интересующегося читателя можно отнести к работам [Социальное исследование: построение и сравнение показателей, 1978; Логика социологического исследования, 1985. Гл. 2; Батыгин, 1981]. Выскажем лишь два коротких замечания.

Во-первых, напомним, что одним из самых известных специалистов по соответствующим вопросам являлся П.Ф.Лазарсфельд [Лазарсфельд, 1972; Батыгин, 1990]. И творчество его, несомненно, должно изучаться каждым социологом, занимающимся эмпирическими исследованиями. Подчеркнем, что Лазарсфельд, глубоко анализируя соотношение наблюдаемого и ненаблюдаемого – ответов респондентов на вопросы анкеты и скрытых факторов, определяющих эти ответы, – разработал соответствующую теорию, сформулированную им на математическом языке и названную латентно-структурным анализом (описание метода можно найти, например, в работах [Моделирование социальных …, 1993; Осипов, Андреев, 1977; Статистические методы …, 1979, с . 249-266; Типология и классификация…, 1982, с. 99-109; Толстова, 1998; McCutcheon, 1987] ). В числе его работ – книга "Математическое мышление в социальных науках" (1954). Название говорит само за себя.

Во-вторых, продолжая сказанное в предыдущей сноске, заметим, что при построении признакового пространства имеет смысл использовать наработки смежных с социологией наук. В качестве одного из предложений, направленных на повышение методологического уровня работы социолога, может служить предложение активного использования разработок, осуществленных в психосемантике по поводу изучения понятия "смысл" и "значение" (об этом мы коротко говорили в [Толстова, 1997]), в частности, использования методов семантического дифференциала и репертуарных решеток [Петренко, 1997].

[7] В связи с тем, что мы в качестве примеров используем какие-то факты, связанные с анкетными опросами (столь популярными у социологов), отметим, что нам очень не хотелось бы, чтобы у читателей сложилось мнение, как будто мы считаем, что анкетные методы – самый хороший способ сбора данных для социолога. Напротив, на наш взгляд, в социологии очень и очень нередки ситуации, когда надо идти другим путем. В данной работе мы не имеем возможности подробно говорить о негативных моментах некоторых часто практикующиеся отечественными социологами подходов к общению с респондентом.

Приведем один пример. Спрашивая респондента о его удовлетворенности своим трудом, и предлагая ему пять вариантов ответа от "совершенно не удовлетворен" до "полностью удовлетворен" (что обычно кодируется цифрами либо от 1 до 5, либо от -2 до +2 и т.д. ), мы предполагаем, что респондент действительно является "носителем" такой удовлетворенности и что он в состоянии выбрать ответ, адекватный его жизненной ситуации. И если один респондент отметил цифру -2, а второй - +1, то первый – носитель меньшего количества положительных эмоций по отношению к работе, чем второй. В действительности же это положение отнюдь не всегда является очевидным: так, разница в ответах может объясняться различием не удовлетворенностей, а манеры поведения (первый – брюзга, а второй всегда по-американски улыбается), понятие удовлетворенности может быть многомерным и т.д. Более подробно об этой проблеме и о возможных подходах к ее решению мы говорим в публикациях, специально посвященных проблеме измерения в социологии, например, в [Толстова, 1998]. Там же осуществляется критика некоторых других традиционных для социологии подходов к измерению.

[8] В данной работе отвлекаемся от глубокого обсуждения проблемы, связанной с анализом того, что есть закономерность развития общества и существуют ли такие закономерности в принципе. В литературе этот вопрос широко обсуждается. См., например [Штомпка, 1996]

[9] Существует даже такая точка зрения, что детерминистских закономерностей вообще не существует. Всё статистично. Так, в [Паниотто, Максименко, 1982] приводится мнение известных ученых о том, что даже законы Кеплера "определяют только средние пути движения планет, от которых последние отклоняются то в ту, то в другую сторону." Мы не хотим здесь обсуждать этот вопрос. Ограничимся констатацией важности статистического подхода для социологии.

[10] В последние годы много говорят о многопарадигмальности в социологии. В качестве основных парадигм выделяют две. В соответствии с первой первичными в развитии общества

являются социальные структуры, детерминирующие поведение отдельного человека (социальный реализм). В соответствии со второй – первичны взаимодействия между отдельными людьми, жизненный мир этих людей, именно он определяет структуру общества в целом (социальный номинализм). При обсуждении соответствующих положений в литературе, к сожалению, имеется много путаницы, нет единства терминов и т.д. Мы не хотим здесь вдаваться в существо вопроса. Отметим лишь, что некоторых недоразумений, на наш взгляд, можно избежать, если разделить все возможные парадигмы на две группы по другому основанию: выделить среди них (в определенной мере – условно) содержательные и методные. Указанные выше парадигмы - частный случай содержательных. Среди методных надо в первую очередь назвать статистическую и системную. Первая парадигма – это та, суть которой является основным предметом рассмотрения в данной работе. В соответствии со второй, мы изучаем рассматриваемый социальный объект (социальную ли группу, отдельного ли индивида – не важно) как систему, придерживаясь соответствующих принципов. Заметим, однако, что названные парадигмы отнюдь не противоречат друг другу. Напротив, они могут эффективно использоваться вместе. См. например, [Сачков, 1999], где рассматриваются статистические системы.

[11] Здесь мы неявно полагаем, что совокупность изучаемых объектов представляет собой некоторую систему. Это означает, в частности, то, что свойства этой совокупности не сводятся к "сумме" свойств отдельных составляющих ее элементов. О том, что общество система, вряд ли в наше время кто-нибудь серьезно сомневается (другое дело, что соответствующие принципы далеко не всегда изучаются и практически используются, хотя библиография по этому вопросу огромна; анализом общества и его составляющих как систем занимались многие выдающиеся исследователи, например, Конт, Спенсер, Дюркгейм, Парк, Парсонс и т.д.).

По всей вероятности, именно системная парадигма активно должна быть использована для изучения современного российского общества (см., например, [Пригожин, 1991]). В последние годы в литературе все чаще высказывается предположение о том, что системная парадигма (при этом чаще всего говорят о синергетическом подходе, соответствующая литература указана в конце п.1.2) может лечь в основу разработки единой социологической теории.

[12] Методы моделирования часто опираются на расчет дифференциальных уравнений, отражающих скорость изменения того или иного процесса, либо на матричную алгебру. Обеспечение потенциальной возможности для будущего социолога читать что-то из огромного пласта литературы по моделированию социальных процессов - одна из причин, почему

традиционный курс высшей математики является необходимой составляющей социологического образования.

[13] Иногда мягкие методы называют нетрадиционными. Однако нам такая терминология представляется сомнительной. Известно, что западная эмпирическая социология в современном понимании этого слова начиналась на стыке XIX и XX веков учеными чикагской школы с активного использования именно мягких методов сбора данных, например, биографического (Парк, Берджесс), неформализованного интервью (Парк, Берджесс, Томас, Знанецкий), анализа писем и официальных документов (Томас, Знанецкий) и т.д. Вероятно, пионерами в области использования мягких методов опроса при решении социологических проблем можно считать русских земских статистиков, проводивших опрос крестьян на деревенских сходах. "Мягкая" сторона земских опросов обеспечивалась, в частности, за счет умения исследователей вызвать заинтересованность опрашиваемого населения, за счет тщательной подготовки интервьюеров (в частности, в ряде областей России интервьюер, прежде, чем приступить к работе, должен был прожить среди крестьян несколько месяцев). Высокие требования, предъявляемые к статистикам-регистраторам, общее представление об их работе как о самоотверженном акте, направленном на улучшение жизни народа, привело к тому, что осуществляемый ими опрос населения оказывается возможным рассматривать, в значительной мере, – как метод включенного наблюдения. За счет проведения опроса на деревенском сходе появлялись в деятельности русских земских статистиков и элементы подходов, которые в наше время называются неформализованным глубинным интервью и методом фокус-групп.

[14] Процесс построения концептуальных моделей, особенно в социологии, является сложным и неоднозначно воспринимаемым разными исследователями. В ряду работ, посвященных соответствующей проблематике, можно выделить разработки, осуществляемые под руководством С.П.Никанорова, касающиеся серьезного изучения процессов концептуального моделирования и концептуального проектирования. Представляется, что эти разработки могут быть полезны для социологии, поскольку они опираются на предложенные авторами способы формализации упомянутых процессов, использующие нетрадиционный для социологических исследований математический аппарат. Это дает возможность избежать многих ошибок (ср. с п.2.1). Об этом см., например, [Никаноров, 1995], а также выпускаемые Ассоциацией концептуального анализа и проектирования научно-практические сборники “Проблемы и решения” и “Подмножество”.

[15] Важно отметить, что причинно-следственные отношения не подлежать формализации. Статистические методы играют огромную роль в их изучении. Однако эти методы могут

подтвердить гипотезу о наличии тех или иных причинно-следственных отношений между рассматриваемыми переменными, заставить исследователей отвергнуть или скорректировать ее, но никогда не могут обеспечить строгое ее доказательство. Яркое подтверждение этого можно найти при использовании методов причинного анализа – с их помощью можно, например, продемонстрировать, что даже очень сильная статистическая связь между двумя переменными может объясняться отнюдь не наличием непосредственной причинной связи между ними, а опосредованной, сложной системой причинных отношений между всеми учитываемыми признаками. Заметим, что эта ситуация как-то перекликается с известными контовским положением о том, что наука должна отвечать на вопрос "как?", а не на вопрос "почему?" См. также п.2.1.2 части II.

[16] Иногда, наряду с рассмотренными нами видами моделей рассматривают также информационную и компьютерную модели изучаемой системы (см., например, [Компьютерное моделирование..., 1994]). В литературе нет однозначного понимания этих терминов; выдвигаемые некоторыми авторами положения, на наш взгляд, весьма спорны (в частности, это можно сказать о названной выше работе). Мы не хотим вступать в дискуссию. Отметим лишь, что совокупность исходных данных (результатов наблюдения) можно назвать информационной моделью изучаемой системы. Компьютерная модель здесь нас не интересует, хотя бывают случаи, что желание повысить качество программы для ЭВМ заставляет исследователя использовать такие элементы формальных алгоритмов, за которыми можно усмотреть наличие вполне определенных и не всегда приемлемых априорных содержательных концепций, касающихся изучаемого социального явления. Последнее обстоятельство, естественно, не может быть проигнорировано социологом.

[17] В литературе по методологии науки обычно принимается утверждение о том, что “те из гипотез следует считать законами, которые при одинаковой их подтвержденности на экспериментальных данных наиболее фальсифицируемы, просты и/или содержат наименьшее число параметров” [Витяев, 1998]. О подтверждаемости и фальсифицируемости мы не говорим (эти понятия здесь интерпретируются в соответствии с пониманием статистической зависимости). Обратим внимание на требование минимизации количества параметров. Об этом требовании говорят очень разные исследователи. Так, оно по существу совпадает с основанными на принципе экономии мышления идеалом чистого описания и понятием истины в махизме. Экономия мышления - это такое описание опыта, которое способно описать короткой формулой огромное количество фактов. Истина – экономная форма описания опыта [Никитина, 1996, с.15] (с последним положением мы не согласны).

[18] Естественно, цели и задачи практически любого научного исследования нельзя свести к трем перечисленным. В литературе, помимо них, выделяются и другие цели: теоретико-познавательная, практически преобразовательная, мировоззренческая, просветительская и т.д.

В работах по методике социологических исследований те качества научного исследования, о которых шла речь, иногда связывают с видом исследования: часто выделяют именно указанные три вида – описательное, объяснительное и предсказательное. Однако надо сказать, что это – неглубокое рассмотрение проблемы. Следовало бы говорить, прежде всего, о классификации самих видов исследований (скажем, по глубине анализа, методу сбору данных, временной продолжительности, виду изучаемых объектов), а уже затем – о выделении видов внутри каждого класса. Например, по глубине анализа можно было бы выделить эмпирическое (прикладное, описательное), теоретическое (фундаментальное, аналитическое, объяснительное, прогнозное), смешанное, пилотажное (зондажное, разведывательное), уточняющее, принципиально новое. По методу сбора данных - выборочное, сплошное, монографическое, сравнительное, опрос (масса видов), анализ документов, наблюдение. По временной продолжительности - однократное, вторичное, лонгитюдное, панельное. По виду изучаемых объектов - изучение отдельных людей и социальных групп, документов, разного вида текстов и т.д.

Кроме того, на практике чаще всего встречаются смешанные виды исследований, в них ставятся сразу несколько целей и т.д.

[19] Понятие эмпирической системы восходит к теории измерений [Суппес, Зинес, 1967]. Однако здесь мы несколько расширяем его. В частности, полагаем, что эта система обладает теми свойствами, которые связаны с априорными предположениями о характере изучаемого явления. О таком расширительном толковании понятия ЭС более подробно идет речь, например, в [Интерпретация и анализ..., 1987.Гл.1; Клигер и др., 1978; Толстова, 1991а, 1998]. Там же рассматривается целый ряд аспектов интерпретации данных, не затрагиваемых в настоящей работе (например, связанных с нашими представлениями об их порождении, о восприятии респондентом предлагаемых ему суждений, объектов и т.д.).

Представляется, что частному случаю рассмотренного нами аспекта обобщения понятия ЭС отвечает представление о вспомогательной теории измерений Блейлока [Blalock, 1982], введенное для учета в процессе измерения гипотез об изучаемых далее связях (напомним, что Блейлок работал в области причинного анализа, а этот метод многомерного анализа предполагает априорное задание системы парных причинных связей между переменными). Примерно те же соображения высказываются Гуттманом в его президентском послании

Психометрическому обществу [Guttman, 1971]. Он говорит о том, что в рамках измерения необходима разработка специальных теоретических конструкций и что теория измерений, в отличие от статистической теории, имеет дело не с выводами из выборки, а с конструированием структурных гипотез. Но Гуттман, на наш взгляд, слишком узко понимает конструированные гипотезы: как и Блейлок, он имеет в виду только структуру связей между переменными.

[20] В наиболее известном фрагменте теории измерений (связанном со строгим определением шкалы) [Суппес,Зинес,1967] в аналогичном случае используется термин "числовая система". Мы расширяем это понятие подобно тому, как расширяем понятие ЭС. О необходимости рассмотрения нечисловых ЭС говорится во многих работах, лежащих в русле теории измерений (не будем здесь указывать библиографические ссылки; их можно найти, например, в [Толстова, 1998]). Сделаем некоторые дополнительные замечания по поводу роли числа в социологии.

"Культ" числа в работах, посвященных проблемам измерения (в разных науках, в том числе и в социологии) связан с тем, что со времен древних шумеров, египтян, греков человеческая цивилизация развивалась именно "под знаменем" числа. Число глубоко вошло в нашу культуру, мы даже не задумываемся о том, что может быть по-другому. А ведь "числовой" характер нашей цивилизации являет собой социальный факт, который, по Дюркгейму, надо рассматривать "как вещь", т.е. взглянуть на него как бы со стороны, "вытащив" себя из привычных концепций (напомним, что к тому же нас призывают этнометодологи и другие сторонники изучения "жизненного мира"). Тогда вполне можно было бы представить, что, если бы древние ученые несколько по-иному взглянули на мир, выделили бы в нем в качестве основополагающих нечисловые конструкции, наша наука могла бы быть другой. В частности, совсем не числовые абстракции должны родиться в голове человека, не предвзято (как на "вещь") смотрящего на человеческие отношения (здесь были бы естественны некоторые соотношения, в наше время рассматриваемые в рамках теории абстрактных алгебр, что, заметим, фактически учитывается при рассмотрении не просто эмпирических и числовых систем, а систем с отношениями; такой подход стал уже классическим, он лежит в основе определения известных типов шкал - номинальной, порядковой, интервальной и т.д.).

А ведь древние греки были не далеки от соответствующих представлений. Так, пифагорейцами было введено понятие гномона (γνωμων – знаток, толкователь; тот, кто знает). Это число или фигура, которая, будучи приложенной к другой фигуре, сохраняет её форму. Сначала гномоном были названы солнечные часы, т.е. прибор, позволявший по линиям, пересекавшим тень от вертикального столбика, разделять беспредельность времени на зримые

части. Впоследствии число стало для пифагорейцев таким гносеологическим гномоном, дававшим возможность различать вещи и тем самым овладевать ими в сознании. Методом гномона растут все живые организмы, что позволяет им сохранять свою индивидуальную форму [Волошинов, 1993, с. 120].

Не будем здесь углубляться в обсуждение вопроса о том, как пифагорейское число помогает различать вещи и при чём тут сохранение формы. Дадим собственную интерпретацию и обобщение "гносеологического гномона".

Итак, констатируем, что, действительно, число часто помогает различать вещи – по тому, сколько чего-то в каждой вещи содержится и по тому, какова пропорция чего-то, отвечающая каждой вещи. Но различие вещей может осуществляться и с помощью каких-либо отвечающих им нечисловых (но, тем не менее – математических; достаточно четкое вычленение конструкций, общих для многих реальных объектов приводит нас к математике в силу самого определения последней) структур. Например, в качестве таких структур могут выступать математические решетки – модели частично упорядоченных множеств, весьма часто встречающихся в социологических исследованиях (для обозначения подобных структур, рассматриваемых как результаты измерения, нами был введен термин "математическая модель структуры эмпирических данных" [Логика социологического ..., 1985, с.104-138; Толстова, 1991а]).

Подобные структуры должны в определённом смысле "сохранять форму" вещи (например, быть гомоморфными выделенным её аспектам; напомним, что, в соответствии с принципами теории измерений, именно понятие гомоморфизма лежит в основе определения шкалы [Суппес и Зинес, 1967; Толстова, 1998]).

Последнее (по порядку перечисления, но не по важности) наше замечание состоит в том, что выделение "формы" вещи в определённом плане аналогично введению понятий, терминов, слов в языке. Это имеет самое непосредственное отношение к определению признаков в социологии.

[21] О прямой и обратной интерпретации такого рода подробно говорится в работе [Интерпретация и анализ ..., 1987]

[22] В подтверждение того, что не всегда простым является соблюдение логической последовательности в рассуждениях (см. наше "во-вторых"), приведем цитату из работы [Рыбников, 1979, с.67] : "...в теоретических изысканиях науки ... цепочки логических суждений ... стали весьма длинными. Опосредованность связей, отражаемых наукой, порождает суждения, состоящие из очень большого числа логических высказываний. При таких условиях неточности

и неоднозначности, допускаемые в определении исходных высказываний и во время “логических ходов” приводят нередко к ошибкам.”

О том же, что круг используемых учеными умозаключений включает в себя рассуждения, не всегда доступные "рядовому" человеку (наше "в-третьих"), свидетельствует следующее своеобразное рассуждение классика американской литературы относительно одного из самых фундаментальных положений теории вероятностей: "…обычного читателя почти невозможно убедить, что при игре в кости двукратное выпадение шестерки делает почти невероятным выпадение ее в третий раз и дает все основания поставить против этого любую сумму. Заурядный интеллект не может этого воспринять, он не может усмотреть, каким образом два броска, принадлежащие уже прошлому, могут повлиять на бросок, существующий еще пока только в будущем.. Возможность выпадения шестерки кажется точно такой же, как и в любом случае – то есть зависящей только от того, как будет брошена кость. И это представляется настолько очевидным, что всякое возражение обычно встречается насмешливой улыбкой, а отнюдь не выслушивается с почтительным вниманием. Суть скрытой тут ошибки – грубейшей ошибки – я не могут объяснить в пределах места, предоставленного мне здесь, а людям, искушенным в философии, никакого объяснения и не потребуется. Тут достаточно будет сказать, что она принадлежит к бесконечному ряду ошибок, которые возникают на пути Разума из-за его склонности искать истины в частностях." [По, 1980. С.228].

[23] Эта проблема очень серьезна. Хотя существует обширная литература, посвященная построению выборки в социологии, все же, наверное, мы не слишком сгустим краски, если скажем, что здесь имеется больше нерешенных проблем, чем решенных. Не будем касаться соответствующих вопросов. Они требуют самостоятельного рассмотрения. Упоминание об этих проблемах нам нужно лишь для того, чтобы даже для неподготовленного читателя стал ясен смысл основных задач, решаемых математической статистикой, - задач, значимость которых для социологии мы должны оценить.

[24] Строго говоря, тому, что в математической статистике называется функцией распределения, отвечает указание для каждого значения “а” случайной величины

вероятности того, что для случайно выбранного респондента отвечающее ему значение этой величины будет *меньше* “а”.

[25] От этого она не становится практически бесполезной. Проведем параллель с такой абстракцией, как прямая линия: ее в природе тоже не существует, однако вряд ли кто-нибудь будет сомневаться в значимости соответствующего понятия для практики.

[26] Даже если мы проведем так называемое сплошное обследование интересующей нас совокупности объектов, как правило, “за бортом” останутся какие-то значения признаков, которые в принципе могли бы служить результатами наблюдения, которые органически вписываются в нашу генеральную совокупность, но которых по чистой случайности в рассматриваемый момент в ней не оказалось. Например, в ней может не оказаться человека, имеющего возраст ровно 20 лет при наличии людей 19-ти лет и 21 года. Вряд ли в таком случае мы будем полагать, что 20-летние люди в принципе должны быть исключены из нашей совокупности.

[27] Термин “статистика” используется в литературе по крайней мере в четырех разных смыслах: как вид деятельности, направленный на получение, обработку и анализ информации, характеризующей количественные закономерности развития общества, во всем их многообразии; как совокупность данных о каком-либо явлении; как отрасль науки, в которой изучаются общие вопросы измерения и анализа массовых количественных отношений и взаимосвязей (в частности, математическая статистика); как обозначение функции от результатов наблюдений. В данном контексте этот термин используется в последнем смысле.

[28] Основные свойства выборочных оценок параметров генеральных распределений сводятся к требования их несмещенности, состоятельности, эффективности. Соответствующие определения можно найти в любой книге по математической статистике. Здесь напомним читателю-социологу только то, что выполнение этих требований повышает вероятность того, что, имея дело лишь с одной-единственной выборкой, мы получим такую оценку интересующего нас параметра, которая “похожа” на генеральное значение последнего. Напомним также, что именно требование несмещенности выборочной оценки дисперсии приводит к тому, что в знаменателе известной формулы стоит не объем выборки, а объем выборки без единицы.

[29] Эти книги представляются полезными для читателя-гуманитария, поскольку в них, на наш взгляд, удачно сочетаются достаточные подробность, строгость и понятность изложения, что не так часто встречается в литературе

[30] Подобные соображения заставляют некоторых авторов вообще отказаться от того, чтобы при сборе и анализе социологических данных использовать какие бы то ни было вероятностные модели порождения данных [Чесноков, 1982,1986]. Несмотря на то, что в настоящей работе мы обсуждаем именно статистические социологические закономерности и полагаем, что они занимают существенное место в деле познания социальных явлений, тем не менее, мы отнюдь не отрицаем целесообразности подхода, разработанного автором указанных

работ. Об этом подходе мы уже говорили в конце п.2.1. Здесь добавим, что кажущаяся эклектичность нашей точки зрения имеет право на существование в той же мере, в какой такое право имеет модное в наше время утверждение о многопарадигмальности социологии [Ядов, 1995].

[31] Здесь представляется уместным отметить, что вряд ли возможно таким образом построить изложение всего материала, касающегося использования в социологии математического аппарата, чтобы он удовлетворял строгим критериям логики, т.е. чтобы каждое вводимое понятие опиралось бы только на уже рассмотренные положения. В частности, это касается сочетания знаний по анализу данных и теории измерений. Казалось бы, сначала надо рассмотреть все, что касается измерения, т.е. обеспечения того фундамента, на котором социолог должен строить свою дальнейшую работу, а уже потом переходить к изложению методов анализа данных, имея в виду, естественно, те данные, которые получены в результате измерения. Но построить курс соответствующим образом оказывается невозможным: говоря об измерении, необходимо говорить о практических способах его осуществления в социологии, а реализация этих способов базируется на ряде положений анализа данных и математической статистики.

Вообще говоря, такой “логический круг” не является случайным. Между измерением и анализом полученных на его основе данных существует определенная связь, носящая весьма принципиальный характер. Истоки этой связи можно проанализировать, если достаточно глубоко изучить роль математического аппарата как средства познания социальных явлений. Об этом идет речь в п. 1.3, п. 2.2, п.5.2 . См. также [Толстова, 1994].

[32] Здесь, правда, необходимо отметить, что вопросы типа описанного не всегда корректно “работают”. Об этом мы уже говорили в сноске [7].

[33] Скажем, в известном пакете SPSS в одном алгоритме классификации CLUSTER предусматривается возможность использования 6-ти способов измерения расстояний между объектами и 7-ми способов - расстояний между классами. Итого - 42 варианта классификации. Каждый, вообще говоря, приведет к своему результату. Что делать “бедному” социологу?

# *Часть 2.*

# ОПИСАТЕЛЬНАЯ СТАТИСТИКА. ИЗМЕРЕНИЕ СВЯЗИ МЕЖДУ НОМИНАЛЬНЫМИ ПРИЗНАКАМИ

Перейдем к подробному рассмотрению конкретных методов анализа данных – методов, позволяющих искать статистические закономерности в "нехорошей" (с точки зрения классической математической статистики) ситуации, специфичной для эмпирического социологического исследования. Наряду с описанием каждого метода, коснемся некоторых методологических принципов их использования из числа тех, которые были рассмотрены в первой части.

Напомним, что основной объект изучения математической статистики – случайная величина – в интересующем нас случае превращается в привычный социологу признак (отвечающий, скажем, какому-либо вопросу анкеты; пол, возраст, удовлетворенность жизнью – примеры признаков); в качестве случайных событий рассматриваются только те, которые состоят в том, что какие-то признаки принимают определенные значения (например, событие может состоять в том, что, взяв анкету, исследователь увидел, что ему "попался" мужчина старше 30 лет, крайне недовольный жизнью); в качестве "хорошей" оценки вероятности того или иного события выступает относительная частота его встречаемости в конкретной изучаемой социологом выборке (мы считаем, что описанное выше событие имеет вероятность 0,15, если доля мужчин с указанными свойствами в изучаемой выборке составляет 15%).

## 1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА.

Как мы отмечали в первой части, социолог практически всегда начинает свою работу с некоторого описания интересующей его совокупности объектов. Для этой цели чаще всего используется расчет частотных распределений (одномерных, двумерных, многомерных), разных показателей среднего уровня значений какого-либо признака, а также индикаторов разброса таких значений. О подобных характеристиках и пойдет речь в данном разделе.

## 1.1. Одномерные частотные распределения.

### 1.1.1. Представление одномерной случайной величины в выборочном социологическом исследовании. Стоящие за ним модели

Итак, в выборочном социологическом исследовании случайная величина предстает перед социологом в виде признака, для каждого значения которого (а таких значений – конечное

количество) известна относительная частота его встречаемости. Эта частота интерпретируется как выборочная оценка соответствующей вероятности (вопрос о правомерности такой трактовки не прост; здесь мы его не рассматриваем; см. п.4.1 части I). Совокупность частот встречаемости всех значений признака, соответственно, трактуется как выборочное представление функции плотности того распределения вероятностей, которое и задает изучаемую случайную величину. Подчеркнем, что пока речь идет об одномерной случайной величине (ниже, переходя к оценке вероятностей встречаемости сочетаний значений разных признаков, мы тем самым перейдем к многомерным случайным величинам).

Пусть, например, вопрос в используемой социологом анкете звучит: "Какова Ваша профессия ?" и сопровождается 5-ю вариантами ответов, закодированных числами от 1 до 5. Тогда частотное распределение - аналог функции плотности - будет иметь, например, вид:

*Таблица 1.*

**Пример одномерной частотной таблицы**

| Значение признака | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Частота встречаемости (%) | 20 | 15 | 25 | 10 | 30 |

Вместо процентов могут фигурировать доли: 20% заменится на 0,2, 15 - на 0,15 и т.д. (в случае такой замены мы получим числа, конечно, в большей степени похожие на вероятности, поскольку величина вероятности, как известно, изменяется от 0 до 1).

То же частотное распределение можно выразить по-другому, в виде диаграммы вида, отраженного на рис. 1 или в виде т.н. полигона распределения, рис.2.

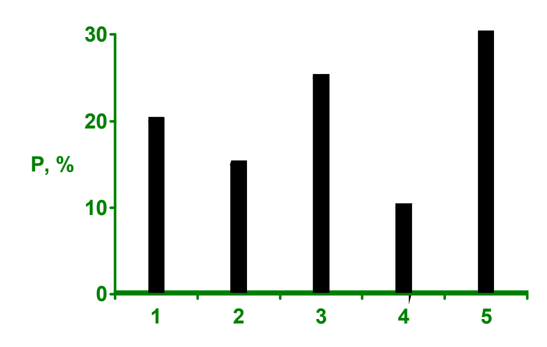



**Рис.1. Диаграмма распределения, рассчитанная на основе таблицы 1.**

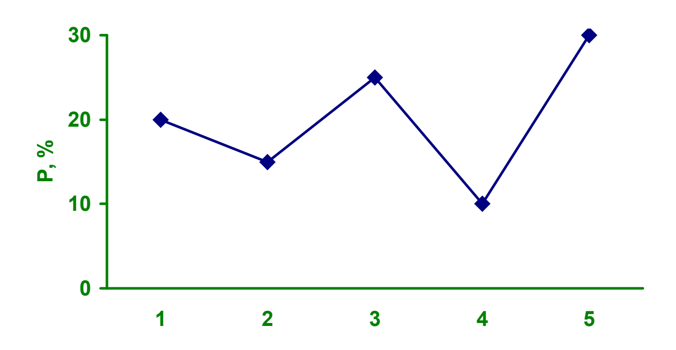



**Рис. 2. Полигон распределения, рассчитанный на основе таблицы 1.**

Подчеркнем, что здесь линии, связывающие отдельные точки, проведены лишь для наглядности, никакой содержательный смысл за ними не стоит (обращаем внимание читателя на то, что ниже ситуация изменится; здесь нельзя говорить об интерпретации линий из-за того, что признак – номинальный).

Казалось бы, что построение частотной таблицы или полигона распределения – дело простое, и говорить не о чем. Однако в социологии это не так. Рассмотрим проблемы, которые возникают при построении одномерных частотных таблиц. Будем учитывать тип шкалы, по которой получаются значения признака, рассмотрим номинальные, порядковые, интервальные шкалы. Однако прежде сделаем некоторое отступление для объяснения того, почему, обосновав во Введении целесообразность ограничиться номинальными данными, мы как будто отступаем от собственных принципов, переходя к шкалам более высокого типа. Дело в том, что продолжая считать номинальные данные основным объектов нашего изучения, мы не можем полностью отвлечься от других шкал. Причин тому несколько.

Во-первых, соответствующие положения фактически задействованы (иногда в неявном виде) почти во всех методах анализа, в том числе и рассчитанных на номинальные данные.

Во-вторых, хотя номинальные данные являются основным предметом изучения социолога, решение большинства задач эмпирической социологии требует “увязки” процесса такого изучения с анализом данных, полученных по шкалам высоких типов. Объясняется это тем, что именно по таким шкалам измеряются столь важные для социолога характеристики респондентов, как возраст респондента, его зарплата и т.д. Поэтому строить курс анализа данных вообще без упоминания методов изучения “числовой” информации представляется нецелесообразным.

В-третьих, хотя в литературе имеется немало работ с описанием методов статистического анализа "числовых" данных, однако при этом не всегда достаточно подробно анализируются многие их аспекты, важные для социолога-практика (например, редко затрагивается проблема разбиения диапазона изменения признака на интервалы или проблема пропущенных значений). Мы постараемся ликвидировать этот пробел хотя бы для наиболее часто используемых социологом методов – вычислении мер средней тенденции и разброса для вероятностных распределений.

Именно с "числовых" шкал мы и начнем более подробное обсуждение специфики построения распределений в социологических задачах. Приводимые ниже рассуждения справедливы для интервальных шкал и шкал более высоких типов.

В социологической практике интервальность шкалы обычно сопрягается с ее *непрерывностью*, т.е. с предположением о том, что в качестве значения интервального признака в принципе может выступить любое действительное число, любая точка числовой оси.

Переходя к описанию выборочного представления функции распределения или функции плотности распределения, прежде всего отметим, что непрерывную кривую в выборочном исследовании нельзя получить никогда. Здесь мы не можем иметь, скажем, линию, похожую на известный "колокол" нормального распределения. Причина ясна: наша выборка конечна. Даже если в генеральной совокупности распределение, к примеру, нормально, а выборка - репрезентативна, мы вместо "колокола" получим лишь некоторое его подобие, составленное, например, из отрезков, соединяющих отдельные точки - полигон распределения (рис. 3). Заменяющая непрерывное распределение ломаная линия может состоять также из "ступенек", в таком случае она называется гистограммой распределения (рис. 4).

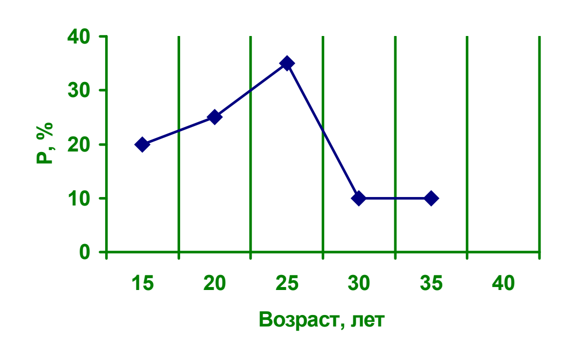



**Рис 3. Полигон плотности распределения непрерывного признака**

От середин отрезков, отложенных на горизонтальной оси, откладываются, соответственно, 20%, 25%, 35%, 10%, 10%

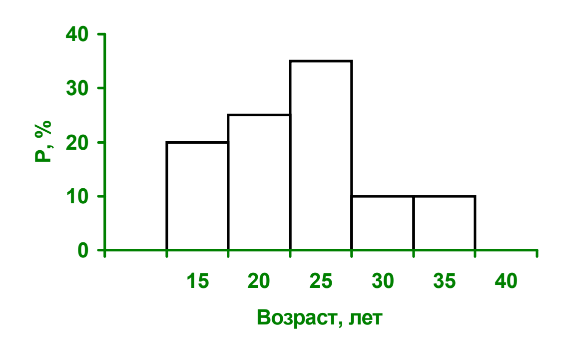



**Рис. 4. Гистограмма плотности распределения непрерывного признака**

В математической статистике доказано, что при больших объемах выборки и достаточно мелком разбиении и гистограмма, и полигон достаточно хорошо приближают функцию плотности распределения (причем полигон делает это несколько лучше) [Ивченко, Медведев, 1992. С.24] (см. также [Тюрин, 1978.С. 8-10; Тюрин, Макаров, 1998. С. 40-41, 319] .

К подробному рассмотрению принципов построения таких “приблизительных” кривых плотностей распределения мы еще вернемся, а пока остановим свое внимание на ситуациях, когда речь идет не о невозможности, а о нецелесообразности стремления к непрерывной кривой.

Для примера рассмотрим признак “возраст респондента”. С одной стороны, без него не обходится практически ни один социолог (вряд ли можно представить себе социологическую задачу, которую имеет смысл решать без учета возраста тех людей, мнения которых изучаются), а, с другой, - на его примере легко демонстрировать некоторые принципиальные положения.

Интересующая нас проблема касается понимания того, чем является та закономерность, которая ищется с помощью того или иного метода анализа данных. Коротко мы же касались этого вопроса в первой части (п.1.4). Продолжим здесь соответствующие рассуждения. Дело в том, что само понятие закономерности предполагает достаточно простую структуру того, что мы закономерностью называем. Слишком дробное описание ситуации мы в силу ограниченности своего мышления (имеется в виду мышление не отдельного человека, а человека вообще) не будем воспринимать как найденную закономерность, как что-то, помогающее нам осмыслить происходящее. Например, мы, всего вероятнее, будем воспринимать сведения о величинах наблюдаемых долей людей с тем или иным возрастом,

выраженные в виде изображенного на рис. 5 фрагмента полигона распределения, как некий бессмысленный набор чисел. А вот если мы сгруппируем соответствующие наблюдения и приведем этот фрагмент к другому виду - виду, изображенному на рис. 6, то нам наверняка станет ясно, что изучаемая совокупность респондентов характеризуется тем, что половину ее составляют люди моложе 20 лет, а людей от 25 до 30 лет в ней вдвое меньше и т.д. Из таких фактов вполне можно сделать содержательные выводы (зависящие, конечно, от того, какую задачу мы решаем). Картину, изображенную на рис. 6, можно назвать закономерностью – пусть весьма примитивной, но

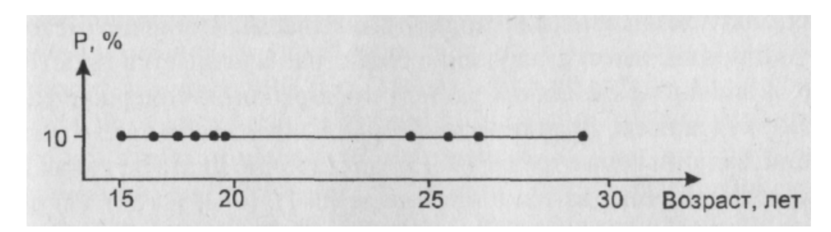



**Рис.5. Полигон распределения по возрасту**

При его построении использовались все наблюденные значения возраста

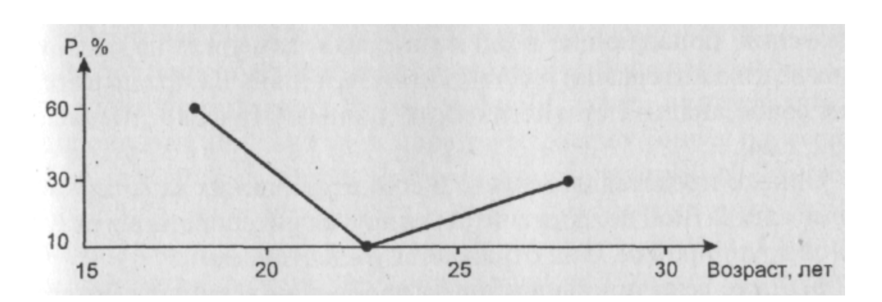



**Рис. 6. Полигон распределения по возрасту**

При его построении объединялись данные, относящиеся к интервалам 15-20 и 25-30

все же закономерностью, поскольку она позволяет нам сформировать какое-то новое представление об изучаемой совокупности респондентов, представление, связанное с описанием совокупности “в среднем”, как целого. Правда, здесь требуется подчеркнуть, что возможна двоякая интерпретация нашего шага.

а) Мы прибегли к определенному “сжатию” информации только потому, что не имели возможности прямо противоположного способа действий: скажем, измерения возраста с точностью до одного месяца и использования репрезентативной выборки в сотни тысяч единиц. Имея возможность сделать это, мы получили бы полигон, неотличимый на глаз от непрерывной кривой.

В таком случае естественно бы было полагать, что мы очень огрубили информацию и ушли дальше от “истинного” распределения, чем находились бы при использовании рис.5. Рассуждая так, мы фактически придерживаемся традиционного для математической статистики восприятия процесса разбиения диапазона изменения признака на интервалы. В соответствии с этим восприятием, указанный подход, называемый обычно методом группировки, имеет следующие свойства: (1) является просто более экономным способом записи информации, содержащейся в выборке (скажем, практически бесполезно знать 10 тысяч наблюдений, заданных на отрезке (0,10), достаточно указать, какая доля наблюдений содержится в интервале (0,1), (0,2) и т.д.), (2) обладает очевидными недостатками, связанными с некоторой неопределенностью в способе построения интервалов и частичной потерей информации при огрублении данных (фактически мы все наблюдения, попадающие в один интервал, заменяем на среднюю точку этого интервала) и (3) используется лишь на предварительном этапе анализа статистических данных [Ивченко, Медведев, 1992. С.24].

Однако представляется, что в социологических задачах часто более адекватной должна считаться другая интерпретация результатов группировки. Она отражается в следующем.

б) Даже если при дальнейшем дроблении величины интервалов распределение респондентов по возрасту будет стремиться к определенному виду, этот вид может вообще не интересовать социолога. Причины – в следующем. Многие “числовые” характеристики людей (в том числе и возраст), чаще всего интересуют социолога не сами по себе (возраст – не как количество оборотов, которые Земля совершила вокруг Солнца за время существования респондента), а лишь как признаки – приборы, как своего рода индикаторы, показатели чего-то непосредственно не измеримого, латентного (например, возраст служит для оценки социальной зрелости опрашиваемого). В таком случае указанное "огрубление" распределения в действительности может служить лишь переходом от признака-прибора к признаку, непосредственно интересующему исследователя (подробнее об этом см. [Клигер и др., 1978; Толстова, 1998]). И наше укрупнение может говорить об интересующем нас распределении больше, чем упомянутый результат дробления. Таким образом, описанная интерпретация частотных распределений – это своеобразное решение одной из проблем социологического измерения.

Итак, при описанной интерпретации имеется налицо, казалось бы, парадоксальная ситуация: если мы хотим получить новое знание с помощью анализа сравнительно небольшого количества наблюденных значений рассматриваемого признака, мы должны “сжать” исходные

данные путем разбиения диапазона изменения значений этого признака на интервалы. За счет потери одной информации, мы приобретаем другую. Здесь тоже хотелось бы сделать определенное обобщение – вычленение какой-либо закономерности из массива “сырых” данных всегда сопряжено с потерей информации. Теряем “сырую” информацию, приобретаем ту, которая содержится в найденной закономерности.

Выбор способа разбиения диапазона изменения признака на интервалы представляет собой проблему, далеко не всегда просто решающуюся. В следующем параграфе рассмотрим ее более подробно. А сейчас приведем пример (заимствованный из [Миркин, 1985. С. 18]), иллюстрирующий, какую огромную роль играет группировка значений признака при анализе данных. При первом чтении книги текст до конца параграфа можно пропустить, поскольку в нем используются положения, рассматриваемые в п.п. 2.1.3 2.3.

Предположим, что мы изучаем связь между двумя признаками: Y, принимающим два значения – 1 и 2, и X, принимающим 4 значения – 1,2,3,4. Предположим, что исходная таблица сопряженности имеет вид (определение таблицы сопряженности дано в п. 1.3 раздела 2; в каждой клетке таблицы указано количество респондентов, обладающих отвечающим этой клетке сочетанием значений рассматриваемых признаков):

**Пример таблицы сопряженности при наличии связи между признаками X и Y**

| X | Y | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1 | 44 | 6 | 50 |
| 2 | 5 | 43 | 48 |
| 3 | 38 | 4 | 42 |
| 4 | 3 | 37 | 40 |
| Итого | 90 | 90 | 180 |

Нетрудно понять, что между X и Y имеется статистическая связь (подробнее о показателях связи см. п. 3 раздела 2). Это можно обосновать, вычислив любой показатель связи, а можно усмотреть и из полуинтуитивных соображений: если бы связи не было, то “внутри” каждого значения признака X респонденты должны были бы поровну распределяться между двумя категориями признака Y (первая строка должна была бы состоять из частот 25 и 25, вторая – 24 и 24, третья – 21 и 21, четвертая – 20 и 20).

Предположим теперь, что мы сгруппировали значения признака X, объединив градации 1 и 2, а также градации 3 и 4 (другими словами, разбили значения признака X на интервалы) .

Получим новую таблицу сопряженности:

**Таблица сопряженности, получающаяся из предыдущей таблицы путем объединения градаций (1 и 2) и (3 и 4) признака X. Связи между X и Y нет**

| X | Y | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1+2 | 49 | 49 | 98 |
| 3+4 | 41 | 41 | 82 |
| Итого | 90 | 90 | 180 |

"Невооруженным" взглядом видно, что никакой зависимости между переделанным признаком X и признаком Y нет. Связь “исчезла”.

Сгруппируем значения признака X по-другому (т.е. по-другому разобьем совокупность этих значений на интервалы): объединим градации 1 и 3, а также градации 2 и 4.

Получим еще одну таблицу сопряженности:

Таблица сопряженности, получающаяся из первой таблицы путем объединения градаций (1 и 3) и (2 и 4) признака X. Связь между X и Y имеется.

| X | Y | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1+3 | 82 | 10 | 92 |
| 2+4 | 8 | 80 | 88 |
| Итого | 90 | 90 | 180 |

Наличие связи представляется очевидным. Связь снова "появилась".

### 1.1.2. Проблема разбиения диапазона изменения признака на интервалы

При определении способа разбиения встает целый ряд взаимосвязанных вопросов: какова величина интервалов? Сколько их? Каково соотношение между ними? И т.д.

Мы не будем подробно рассматривать эти вопросы. Лишь коротко заметим, что их решение в первую очередь должно опираться на содержание задачи. Так, при изучении типов личности, вполне возможно, что нас удовлетворит разбиение всех возрастов от 15 до 100 лет на равные интервалы: (15-20), (20-25), (25-30) и т.д. Если же одной из решаемых нами задач будет изучение выбора молодежью жизненного пути, то мы, вероятно отдельно рассмотрим интервалы (15-17), поскольку в 17 лет человек кончает школу; (17-18), поскольку в 18 лет юношей забирают в армию; (18-22), поскольку в 22 года большинство поступивших после

школы в институт получают дипломы о высшем образовании и т.д. Если нас интересует лишь производственная деятельность людей, то всех лиц старше 60 лет мы будем считать одинаковыми по возрасту (в анкете одним из вариантов ответа на вопрос о возрасте будет вариант “старше 60”). Если нас будут интересовать какие-то аспекты геронтологии, то, возможно мы выделим интервалы (70-72), (72-74) и т.д. [Пасхавер, 1972; Сиськов, 1971]

Конечно, какую-то роль при выборе интервалов разбиения может сыграть желание исследователя иметь возможность сравнивать свои результаты с результатами других социологов - в таком случае способы разбиения диапазонов изменения тех признаков, по которым совокупности сравниваются, должны быть одинаковыми. В свое время были выдвинуты предложения по унификации разбиения на интервалы диапазонов тех признаков, которые обычно входят в стандартную “паспортичку” анкеты. Однако это не прижилось, поскольку все же разные задачи диктуют разные разбиения [Петренко, Ярошенко, 1979].

Существуют и математические методы, помогающие разбить диапазон изменения признака на интервалы [Орлов, 1977]. Однако при этом речь идет о достаточно тонких и сложных моделях того, что происходит в сознании респондента, дающего нам информацию. Здесь мы их рассматривать не будем.

Разбив на интервалы, мы ставим другие вопросы. Рассмотрим наиболее часто встающие.

*К какому интервалу относить объект, для которого значение рассматриваемого признака лежит на “стыке” двух интервалов?* Ответом на него обычно служит соглашение: скажем, все “стыки” считать принадлежащими правому интервалу (используя известные математические обозначения, можно сказать, например, что при разбиении диапазона изменения возраста на равные интервалы по 5 лет, мы в действительности будем рассматривать полуинтервалы: [15, 20), [20, 25) и т.д. Последним полуинтервалом может быть, например, [60, 65). Заметим, что фактически используемая нами при этом модель (мы уже неоднократно подчеркивали, что какая-то модель всегда стоит за любым, даже самым простым, математическим методом, и что для социолога раскрытие смысла подобных моделей является первоочередной задачей) изучаемого явления может привести к неоправданному (хотя вряд ли большому, особенно для многочисленной выборки) сдвигу массива данных вправо. Это скажется, например, при расчете мер средней тенденции (их определение см. ниже).

*Как в только что описанной ситуации поступать с правым концом самого правого интервала?* Прибегая к только что приведенному примеру, переформулируем вопрос: что делать с возрастом 25 лет? Ответы могут быть разными: например, вместо полуинтервала

[60,65) использовать отрезок [60,65]; ввести дополнительный полуинтервал [65,70). При достаточно репрезентативной выборке принятие любого из них приведет примерно к одному и тому же результату (точнее, результаты не будут статистически значимо отличаться друг от друга).

При построении полигонов и гистограмм встают свои вопросы.

*От какой точки интервала проводить вертикаль, на которой будет откладываться величина процента при построении полигона?* На этот вопрос мы ответили в работе [Толстова, 1998] (см. также Приложение 1). Там соответствующая ситуация рассмотрена очень подробно. Здесь же лишь отметим, что вертикаль может начинаться в любой точке интервала (хотя на практике из иллюстративных соображений чаще всего используют его середину).

Конечно, при выборе разных точек, в процессе дальнейшего анализа данных, вообще говоря, будут получаться разные результаты. Однако если считать, что мы работаем в рамках интервальной шкалы, то соответствующее различие будет именно таким, которое с точки зрения теории измерений для этой шкалы вполне допустимо.

*Чем отличаются друг от друга модели, которые мы фактически используем, строя, с одной стороны, - полигон, а, с другой, - гистограмму распределения?*

В обоих случаях мы в процессе построения закономерности (коей является частотное распределение) теряем информацию о том, каким образом распределены объекты внутри каждого интервала, и восполняем эту потерю путем введения модельных предположений об этом распределении. Обычно считают, что полигон отвечает кусочно-линейной плотности распределения. При использовании же гистограммы полагают, что объекты равномерно распределены внутри каждого интервала.

Напомним, что в соответствии с известными положениями теории вероятностей, площадь фигуры, лежащей под кривой функции плотности над каким-либо интервалом равна вероятности попадания объекта в этот интервал. Особенное внимание ниже будет обращено на то, как это свойство проявляется в случае гистограммы (здесь оно превращается в то обстоятельство, что вероятность попадания значения признака на тот или иной отрезок равна площади соответствующего отрезку прямоугольника гистограммы), поскольку площади прямоугольников легко вычисляются.

*Как строить гистограмму с неравными интервалами?*

Способ построения такой гистограммы опирается на только что сформулированное положение о площадях составляющих гистограмму прямоугольников. На примере опишем соответствующий алгоритм.

Предположим, что частотная таблица, на базе которой мы хотим построить гистограмму, отвечающую распределению нашей совокупности респондентов по возрасту, имеет вид, отраженный в таблице 2. .

*Таблица 2*

**Частотное распределение респондентов по возрасту**

| Интервал изменения возраста | [15 - 20) | [20 - 50) | [50 - 55) | [55 - 80) |
|---|---|---|---|---|
| Количество респондентов, попавших в интервал | 80 | 90 | 20 | 10 |

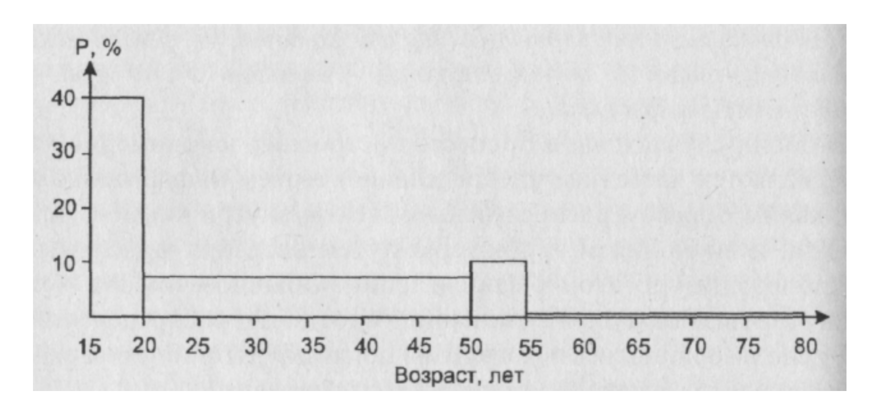



**Рис. 7 . Гистограмма, построенная на основе частотной таблицы 2**

Подчеркнем, что предлагаемое разбиение на интервалы представляется нам разумным для некоторых задач - скажем, в том случае, если мы особенно интересуемся категориями женщин, с одной стороны, думающих о вступлении в фазу трудовой деятельности и вступающих в нее (15 - 20 лет) и, с другой стороны, - собирающихся покинуть эту фазу (50-55 лет) (заметим, что людей старше 80-ти лет в нашей совокупности нет).

Итак, алгоритм состоит в следующем. Выбираем какой-то интервал диапазона изменения возраста за единицу и считаем, что на нем высота столбца гистограммы равна проценту людей, попавших в этот интервал. Для гистограммы, изображенной на рис. 7 - это интервалы [15 - 20) и [50 - 55). Другими словами, мы выбрали за единицу интервал длиной в 5 лет. Для интервалов,

имеющих другую длину, высоту столбца гистограммы будем полагать равной результату деления величины процента попавших в него людей на длину интервала. Так, интервал [50 - 55) имеет длину в 6 наших единиц. В него попали 45% респондентов. Поделим 45 на 6 . Получится 7,5%. Именно такой высоты столбец и будет отвечать рассматриваемому интервалу. Так же поступим с интервалом [55 - 80). В него попало 5% респондентов, а длина его равна 5 единицам. Значит, высота соответствующего столбца равна 50:5 = 1 %.

Нетрудно проверить, что при описанном подходе площадь каждого столбца будет равной проценту респондентов, возраст которых попал в интервал, лежащий в его основании.

Социологу необходимо приучить себя правильно интерпретировать гистограмму и сразу, в результате беглого взгляда на нее, оценивать содержательную суть представленного ею распределения: эта оценка должна базироваться на анализе не высоты столбцов, а их площади! (Роль беглой визуальной оценки графических данных в процессе формирования научных взглядов на изучаемый предмет, анализируется наукой. Соответственно, изучаются разные способы визуализации данных с точки зрения наиболее эффективного воздействия на сознание исследователя, наиболее адекватного улавливания им сути отраженных в “картинках” явлений. Об этом см., например (Плотинский, 1994)).

Именно при описанном подходе к построению гистограммы ее можно считать выборочным представлением того, что в математической статистике называется функцией плотности распределения. Только в этом случае гистограммы, представляющие, скажем, функцию плотности нормального распределения, будут в совокупности по своей форме напоминать известную форму "колокола" и при увеличении дробности интервалов все больше приближаться к идеальной “гладкой” кривой соответствующего вида.

### 1.1.3.Кумулята

Выборочным представлением собственно функции распределения (а не плотности) случайной величины, “стоящей” за рассматриваемым признаком, служит т.н. кумулята распределения, или график накопленных частот. Она обычно представляется в виде полигона, каждая вершина которого отвечает относительной частоте того, что признак принимает значение, не превышающее того, над которым эта вершина находится. Нетрудно понять, что кумулята получается из описанного выше полигона распределения путем последовательного

суммирования определяющих его частот. Так, полигону, изображенному на рис. 6, будет отвечать следующая кумулята (рис.8):

Так, полуинтервалу (25, 30] будет отвечать частота 80%, складывающаяся из отраженных на рис. 3 частот, соответствующих полуинтервалам (15, 20], (20, 25] и (25, 30]. Выборочное представление функции распределения может быть задано и в виде гистограммы (рис. 9).

Теперь вспомним, что непрерывные интервальные шкалы - не самые важные для социолога виды шкал (даже возраст социологом часто рассматривается как номинальная или порядковая переменная: выделяются классы работающих и пенсионеров, молодежи и людей более старших возрастов, репродуктивный возраст и нерепродуктивный и т.д.). Перейдем к рассмотрению *номинального и порядкового* уровней измерения. Шкалы соответствующих типов в социологии обычно бывают дискретными: в анкете используется конечный набор значений (например, удовлетворенность работой может измеряться по семибалльной порядковой шкале; для измерения профессии можно использовать

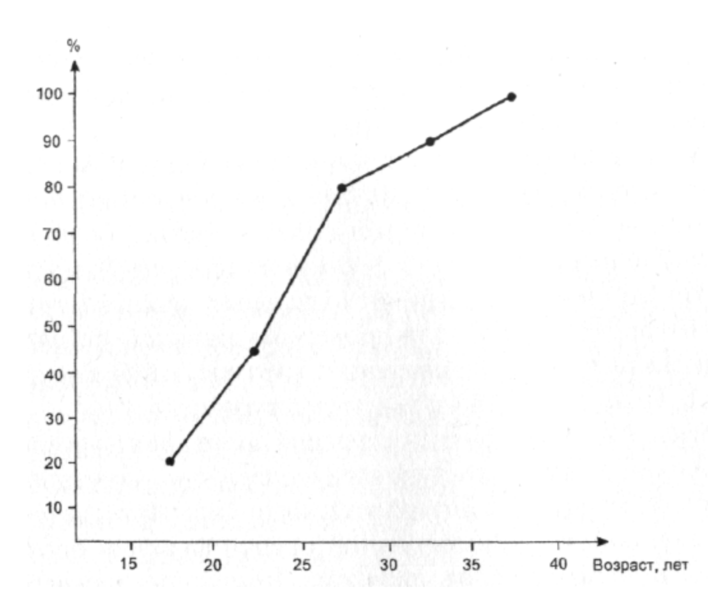



**Рис. 8. Кумулята распределения, отвечающего выборочной функции плотности, изображенной на рисунке 3**

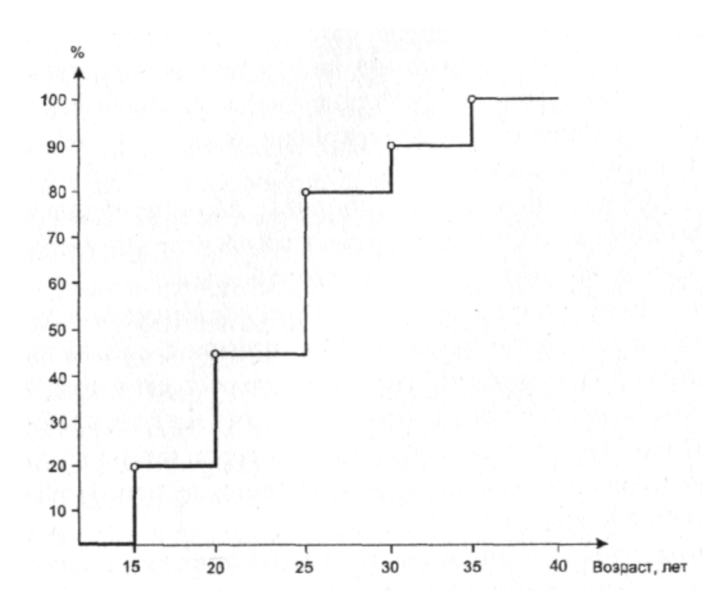



**Рис. 9. Кумулята распределения, отраженного на рисунках 3 и 8, заданная в виде гистограммы**

номинальную шкалу, определяемую, скажем, 30-ю конкретными наименованиями), и встает вопрос о том, как здесь строить полигоны, гистограммы, кумуляты.

Сразу отметим, что говорить о кумуляте для номинальной шкалы в принципе невозможно, поскольку для значений признака, полученных по этой шкале, теряет смысл понятие “больше” или “меньше”. Полигон, как мы уже говорили (см. рис.2), построить можно. Но отрезки, связывающие отдельные точки, мы никак не можем интерпретировать. Они проведены лишь для наглядности и график на рис.2 эквивалентен картине, изображенной на рис. 1. То же можно сказать и о гистограмме.

Относительно специфики построения полигонов и гистограмм для порядковых шкал заметим следующее. Кумуляту для таких шкал строить можно. Но интерпретация полигонов и гистограмм (и для кумуляты, и для выборочной оценки функции плотности распределения) может быть двоякой. Поясним на примере рассмотрения функции плотности.

Возможны два варианта интерпретации результатов измерения по порядковой шкале.

1) Полагаем, что в принципе наш признак непрерывен, а наблюдаемая дискретность (наблюдаемая совокупность значений любого признака всегда дискретна хотя бы в силу своей конечности) объясняется

- либо только конечностью выборки, а в принципе мы можем получить в качестве наблюдаемого значения любое действительное число рассматриваемого отрезка числовой оси;

- либо (что обычно более отвечает реальности) тем, что мы не умеем достаточно точно измерять наш признак; рассматриваем лишь несколько его уровней; измерение же состоит в том, чтобы каждый измеряемый объект отнести к одному из этих уровней.

2) Считаем, что признак дискретен по своей природе, т.е. что для него не имеют смысла числа, лежащие между используемыми шкальными значениями.

В первом случае мы вполне можем интерпретировать полигон и гистограмму так же, как это делали для интервального признака. Во втором же случае построение и того, и другого рассматривается как чисто иллюстративный прием - так же, как это имело место для номинального признака.

### 1.1.4. Проблема пропущенных значений

Социолог постоянно сталкивается с ситуацией, когда значительная часть респондентов не дает ответа на какие-то вопросы анкеты. Если для “исправления” этого положения идти по наиболее простому пути - выбросить анкеты, содержащие хотя бы один пропуск, то мы почти наверняка останемся без репрезентативной выборки, поскольку в ее составе останется слишком мало объектов. Об этом свидетельствует практика социологических исследований.

Неразумно просто исключать из рассмотрения упомянутые анкеты и еще по одной причине. Скажем, зачем нам выбрасывать анкету с неотмеченным возрастом, если мы изучаем связь между доходом респондента и тем, за кого он голосовал на прошлых выборах? Вероятно, имеет смысл, рассчитывая любую статистику, выбрасывать именно те анкеты, в которых отсутствуют сведения, необходимые для такого расчета. Но и здесь мы рискуем отбросить слишком много анкет. Кроме того, у всякого исследователя может возникнуть сожаление о том, что, отбрасывая анкету из-за отсутствия в ней ответа на один из вопросов, он тем самым лишается возможности использовать всю, может быть весьма объемную и полезную информацию, содержащуюся в этой анкете. На помощь в таком случае может придти иной вариант решения проблемы - искусственное заполнение пропусков.

Известно много способов, позволяющих это сделать [Алгоритмы..., 1984; Вапник, 1979; Загоруйко, 1979, с.105-118; Лакутин, 1982; Лбов, 1981, с.38-41,52-55; Литтл,Рубин,1991]. Мы не будем их подробно рассматривать. Отметим лишь следующее немаловажное для социолога обстоятельство.

За каждым методом заполнения пропусков стоит своя модель массива пропущенных данных, свое представление о том, какие именно респонденты допускают пропуски. Применяя тот или иной алгоритм заполнения пропусков, исследователь фактически пользуется заложенной в этом алгоритме моделью, даже если он себе и не дает отчета в этом. Приведем примеры.

Один из самых распространенных способов - заполнение пропуска средним значением рассматриваемого признака (как мы увидим в п.1.2, выбор среднего должен быть согласован с типом используемых шкал). И исследователь должен понимать, что, поступая так, он рискует придать данным более ровный, "серый" характер, чем это имеет место в действительности. Можно поступать по-другому: проанализировать распределение признака для тех респондентов, которые ответили на соответствующий вопрос, и заполнять пропуски таким образом, чтобы получающееся в результате распределение имело тот же характер (этот способ отвечает рассматриваемому в п.2.3.2.3 пропорциональному прогнозу). Этот подход можно улучшать, осуществляя такую операцию не для всех пропущенных данных сразу. К примеру, предположим, что нам надо заполнить пропуски возраста. Распределение по возрасту мужчин может отличаться от аналогичного распределения женщин (имеем в виду людей, ответивших на соответствующий вопрос). Тогда имеет смысл, отобрав мужчин и определив для них вид распределения, далее именно этот вид моделировать при заполнении пропусков, сделанных мужчинами. Затем то же следует проделать для женщин.

В заключение лишь отметим, что существуют интересные работы, посвященные содержательному изучению того, кто именно не отвечает на определенные вопросы, и высказываются гипотезы о том, почему это делается [Клюшина, 1990; Федоров, 1982].

### 1.2. Меры средней тенденции и отвечающие им модели

Итак, мы получили частотное распределение значений рассматриваемого признака, т.е. выборочное представление изучаемой одномерной случайной величины. Конечно, анализ этого распределения может много дать социологу. Именно с расчета таких распределений для всех рассматриваемых признаков (так называемых "линеек") он обычно и начинает анализ данных. Каждое распределение представляет собой своеобразное описание изучаемой совокупности объектов (респондентов). Такие описания позволяют исследователю лучше сориентироваться в проблематике, скорректировать перечень проверяемых гипотез, уточнить априорные

представления об объекте и предмете исследования. Но этим анализ каждого одномерного распределения обычно не ограничивается.

Оказывается, что даже для одномерных случайных величин можно найти целый ряд статистических закономерностей. Конечно здесь они довольно примитивны (скажем, мы не можем говорить о связях между переменными), но все же это - статистические закономерности. В первую очередь мы имеем в виду так называемые меры средней тенденции, среди которых (в математической статистике известно бесконечное количество таких мер, им посвящена довольно обширная литература, см., например, [Джини, 1970]). в социологии наиболее часто используются математическое ожидание, мода и квантили (наиболее употребительным квантилем является медиана). Их мы и рассмотрим, полагая, что необходимость использования этих мер социологом очевидна. Подчеркнем лишь, что каждая из этих мер – некоторое значение (единственное!) рассматриваемого признака, которое должно характеризовать, как бы подменять, всю нашу совокупность. И социолог должен проявлять повышенное внимание к тому, чтобы с содержательной точки зрения такая подмена была оправданной.

Напомним, что названные средние являются параметрами распределения вероятностей. Не будем давать их строгого определения для генеральной совокупности. Опишем лишь то, как они измеряются для выборки. Говоря более грамотно, мы покажем, каковы выборочные точечные оценки указанных параметров, или, что то же самое, опишем способы расчета отвечающих этим параметрам выборочных статистик. (Напомним, что выборочные оценки параметров распределения делятся на точечные, когда для выборочных данных находится одно значение, служащее оценкой генерального параметра, и интервальные, когда на базе выборочной точечной оценки параметра строится так называемый доверительный интервал. Определенная на выборке переменная, значениями которой служат точечные оценки какого-либо параметра, называется статистикой, отвечающей этому параметру. Соответствующий материал обычно изучается в курсе математической статистики; см. также [Гласс, Стэнли, 1976; Статистические методы ..., 1979] .)

Все описываемые ниже меры средней тенденции являются "хорошими" выборочными точечными оценками генеральных параметров (напомним, что "хорошей" оценкой в математической статистике называются оценки, являющиеся несмещенными, состоятельными, эффективными; не будем напоминать, что это такое; отметим только, что выполнение указанных свойств

дает исследователю возможность с наибольшей вероятностью избежать сильного отклонения наблюденного значения статистики от соответствующего генерального параметра).

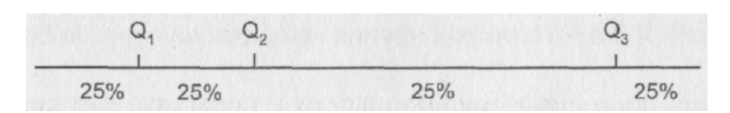
Пусть $x_1$, $x_2$, ..., $x_N$ – выборочные значения рассматриваемого признака (N – объем выборки). Статистикой, отвечающей математическому ожиданию (дающей “хорошие”. точечные выборочные оценки этого параметра; это также – материал курса математической статистики) является знакомое всем *среднее арифметическое* значение признака:

$$\overline{x} = \frac{(x_1 + x_2 + ... + x_N)}{N}$$

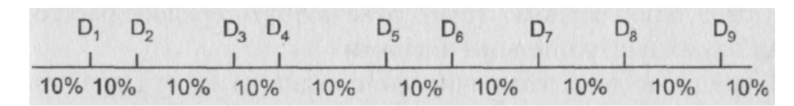
Среднее арифметическое значение признака, вычисленное для какой-либо группы респондентов, чаще всего интерпретируется как значение для наиболее типичного для этой группы человека, это среднее значение как бы служит "олицетворением" этой группы (по качеству, связанному с рассматриваемым признаком). Однако бывают случаи, когда подобная интерпретация среднего арифметического несостоятельна. Ниже мы рассмотрим некоторые из них.

Напомним, что *квантиль* – это такое значение признака q, которое делит диапазон его изменения на две части так, чтобы отношение числа элементов выборки, имеющих значение признака, меньшее q, к числу элементов, имеющих значение признака, большее q, было равно заранее заданной величине. Среди всех возможных квантилей обычно выделяют определенные семейства. Квантили одного семейства делят диапазон изменения признака на заданное число равнонаполненных частей. Семейство определяется тем, сколько частей получается. Наиболее популярными квантилями являются *квартили*, разбивающие диапазон изменения признака на 4 равнонаполненные части; *децили* - на 10 равнонаполненных частей; *процентили* – на 100 частей. Символически эти определения можно изобразить следующим образом.

Квартили:




Децили:




Процентили:

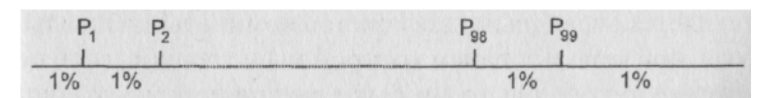


**Рис. 10. Иллюстрация сущности наиболее употребительных квантилей.**

Величина процента, указанная под интервалом означает долю объектов выборки, попавших в этот интервал.

Разного рода квантилями социолог пользуется очень часто. Нередко они упоминаются в средствах массовой информации (однако при этом сами термины "квантиль", "квартиль" и т.д. при этом не используются). Так, в газетах пишут о том, что, например, 10% наиболее богатых "россиян" имеют месячный доход свыше 100 тысяч рублей, а 10% наиболее бедных – ниже 300 рублей. Ясно, что 100 тысяч рублей – это девятый дециль $D_9$, а 300 рублей – это первый дециль $D_1$.

*Медианой* называется $Me = Q_2 = D_5 = P_{50}$.

Нетрудно видеть, что так определенная выборочная медиана – это значение рассматриваемого признака, которое делит отвечающий этому признаку *вариационный ряд* (т.е. последовательность значений признака, расположенных в порядке их возрастания) пополам. Иначе говоря, медиана обладает тем свойством, что половина всех выборочных значений признака меньше нее, а половина – больше. "Правомочность" медианы в качестве представителя анализируемой группы респондентов представляется очевидной. Для того, чтобы это почувствовать, достаточно "взглянуть", скажем, на две группы, в одной из которых медиана признака "доход" равна 500 рублей, а в другой – 5000 рублей. Ясно, что вторая группа "в среднем" гораздо богаче первой.

Обычно, построив вариационный ряд, полагают, что при нечетном числе элементов в выборке медиана равна центральному члену ряда, а при четном – точке, отвечающей середине расстояния между двумя центральными членами.

Нетрудно видеть, что вычисление медианы имеет смысл только для порядкового признака (и, конечно, для интервального, поскольку любая интервальная шкала является порядковой). Это представляется очевидным: для “чисто” номинальной шкалы (т.е. для такой, при использовании которой мы не ставим своей целью отображение какого бы то ни было эмпирического отношения порядка в числовое) само выражение “объект обладает значением признака, меньшим, чем медиана” становится бессмысленным. Понятия “больше” или “меньше” в этой ситуации не существуют

В случае же, когда медиана вычисляется как середина между двумя шкальными значениями, мы делаем фактически еще одно предположение – о том, что наш порядковый

признак в принципе может принимать значения, лежащие между используемыми пунктами шкалы.

Можно рассчитывать медиану и с помощью построения кумуляты. Это также опирается на предположение о непрерывности рассматриваемого признака. Более того, здесь работает еще одно модельное предположение: объекты внутри каждого интервала распределены равномерно. Подчеркнем, что этот пример хорошо иллюстрирует то, что за каждым математическим методом, даже самым простым, стоит своя модель изучаемого явления. В данном случае - модель понимания средней тенденции. Разбив диапазон изменения признака на интервалы и простроив полигон плотности распределения, мы потеряли информацию о том, как в действительности расположены объекты внутри каждого интервала, и заменили эту информацию модельным предположением, состоящим в том, что соответствующее распределение равномерно.

То, как находятся квантили с помощью кумуляты, подробно описывается, например, в [Паниотто, Максименко, 1982; Толстова, 1998; Ядов, 1998]. Мы не будем на этом подробно останавливаться. Надеемся, что суть подхода станет ясной из рис. 11.

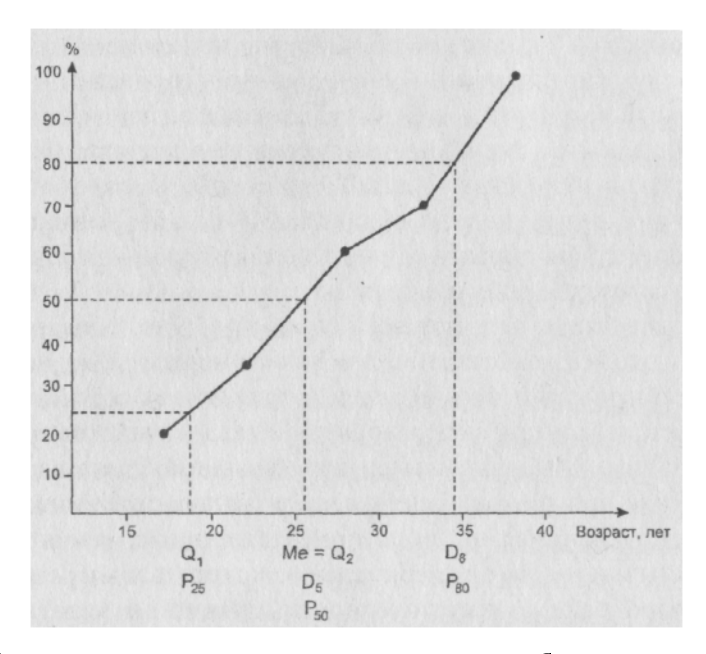



**Рис. 11. Иллюстрация одного из возможных способов расчета квантилей**

Эквивалентным этому подходу является расчет квантилей по формулам, приведенным в [Рабочая книга ..., 1983. С. 161]. Более подробно о разных способах расчета медианы и о сути используемых при этом моделей см. Приложение 1 (на наш взгляд, рассмотрение соображений,

описанных в этом Приложении, может способствовать лучшему пониманию, что такое модель, заложенная в методе).

*Модой* называется наиболее часто встречающееся значение признака. Нахождение моды обычно не представляет трудностей. Ясно, что ее можно рассчитывать для признаков, измеренных по шкалам любых рассматриваемых нами типов. (Иногда моду предлагается рассчитывать по определенной формуле [Рабочая книга ..., 1983. С.162]. Но это сопряжено с довольно сильными модельными предположениями; в частности, признак должен быть порядковым и непрерывным).

Надеемся, что читателю ясно, почему моду относят к мерам средней тенденции. Приведем пример. Сравнивая, скажем, распределение по профессиям, рассчитанные для двух регионов – Ивановской и Тюменской области, мы можем придти, например, к выводу, что в первой наиболее распространенная профессия – ткачиха, а во второй – нефтяник. Этот вывод означает, что ткачиха – модальное значение профессии для жителей Ивановской области, а нефтяник – для Тюменской. И соответствующее первичное описание этих областей, т.е. как бы условное отождествление первой области с ткачеством, а второй – с добычей нефти, является вполне естественным.

Подчеркнем, что, при всей своей простоте, описанные статистики – это все же статистические закономерности, и при их расчете и интерпретации возникает множество тех же методических проблем, что и при использовании сложных многомерных методов анализа. Мы не можем уделить таким проблемам достаточное внимание при рассмотрении всех затрагиваемых ниже методов. Коротко коснемся их лишь применительно к тем простейшим статистическим закономерностям, о которых идет речь в настоящем параграфе. А именно, обратим внимание читателя на следующие, не всегда замечаемые методические аспекты использования мер средней тенденции, пытаясь по возможности обобщить соответствующие положения на ситуации, возникающие при изучении статистических закономерностей произвольного вида.

Как мы уже отметили, любая средняя – это *параметр распределения* соответствующей случайной величины (либо статистика, вычисленная для выборочного частотного распределения рассматриваемого признака). И здесь мы сталкиваемся с общим положением - все известные методы нахождения статистических закономерностей являются методами расчета некоторых параметров рассматриваемых распределений (не обязательно одномерных), любая закономерность может быть выражена через ту или иную совокупность параметров. И для всех таких параметров встает задача их точечного и интервального оценивания. Для средних величин

способы решения этой задачи известны [Гласс, Стэнли, 1976; Гмурман, 1998а; Калинина, Панкин, 1998; Статистические методы ..., 1979]. Однако для многих интересующих социолога параметров не разработана та теоретическая основа, которая дает возможность построения интервала. В таких случаях социолог, вообще говоря, лишается возможности переносить результаты с выборки на генеральную совокупность. Правда, как мы уже отмечали в п.4.1 части I, современная наука предоставляет некоторый способ преодоления этой трудности – использование специальным образом организованной процедуры моделирования большого числа выборок на ЭВМ, наблюдение получающихся при этом распределений рассматриваемых статистик (для каждой выборки - свое значение статистики), вычисление параметров этих распределений и построение на этой основе требующихся доверительных интервалов.

Далее, любая статистическая закономерность – это своего рода *сжатие исходных данных*. Это ярко видно на примере средних величин. Так, при использовании среднего арифметического мы вместо набора, скажем, из 1000 значений возрастов мы получили одно число – 32,4, средний возраст респондентов рассматриваемой совокупности. Совокупность из тысячи чисел сжата в одно число.

Указанное сжатие означает *потерю информации.* С такой потерей связано нахождение любой закономерности (коротко об этом уже шла речь в п.1.4 части I). Анализируя данные, мы всегда сталкиваемся с парадоксом: только потеряв определенную информацию, мы можем приобрести новое знание (содержащееся в найденной закономерности). И интерпретируя найденное статистическое соотношение, постоянно надо давать себе отчет в том, что мы теряем. Так, пользуясь упомянутым выше средним значением, мы как бы забываем про то, что в нашей совокупности могут находиться люди весьма различного возраста. Она для нас начинает ассоциироваться с возрастом 32,4 года, мы как бы полагаем, что именно такой возраст имеет наиболее типичный представитель совокупности. А это может не отвечать действительности.

Следующее обстоятельство касается того, что любая статистическая закономерность имеет смысл лишь при определенной *однородности* той совокупности объектов, для которой эта закономерность рассчитывается. Понятие однородности сложно и многогранно [Толстова,1991а]. В нем имеются аспекты, как не зависящие от того, какую закономерность мы ищем, так и “привязанные” к конкретному методу анализа данных. И отнюдь не для всех важных для социолога методов эти аспекты изучены. Но средним в этом смысле “повезло”. В названной выше работе приведен перечень публикаций, в которых анализируется проблема однородности для среднего арифметического. Интуитивно ясно, о чем здесь идет речь: нельзя

считать среднюю температуру по больнице и на этой основе сравнивать работу разных медицинских учреждений. Нельзя считать среднюю зарплату по какому-либо региону, если различие между высокооплачиваемыми и низкооплачиваемыми людьми слишком велика. В таком случае средняя зарплата не будет информативна. И на ее основе нельзя будет сравнивать, скажем, обеспеченность населения двух регионов.

Как мы отмечали в п.4.3 части I, одним из основных свойств социологических данных, обусловливающих специфические моменты в использовании социологами математической статистики, является то, что эти данные зачастую бывают получены по шкалам низких типов, из которых мы рассматриваем номинальные и порядковые. Метод анализа данных необходимо сопрягать с типом используемых шкал. Результаты применения метода должны быть инвариантными относительно применения к исходным данным допустимых преобразований тех шкал, по которым эти данные получены. Это свойство метода в работе [Толстова, 1998] называется его *формальной адекватностью*.

В свете этого можно сказать, что моду можно вычислять для шкал любых типов, начиная с номинального – объект, обладающий модальным значением, не будет изменяться при любом взаимно-однозначном преобразовании исходных шкальных значений (как известно, эти преобразования являются допустимыми для номинальных шкал). Значит, любые выводы, полученные на основе анализа мод, будут удовлетворять сформулированному выше свойству инвариантности.

Для того, чтобы имел смысл расчет медианы и других квантилей, шкала, как мы уже упоминали, должна быть по крайней мере порядковой. Легко показать, что все выводы на базе анализа квантилей останутся без изменения, если к исходным данным применить монотонно возрастающее преобразование (допустимое преобразование порядковых шкал).

Нетрудно понять, что среднее арифметическое неявно предполагает использование шкалы, отвечающей по крайней мере интервальному уровню измерения. Действительно, среднее арифметическое – это такое значение признака, для которого сумма расстояний от него до объектов, имеющих большее значение, равна сумме расстояний до объектов, имеющих меньшее значение. Это легко вытекает из соотношения:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_N - \bar{x}) = 0$$

В этом – суть рассматриваемой статистики. Стало быть, эта самая суть требует осмысленности соотношений между расстояниями от одних значений признака до других.

Перейдем к рассмотрению свойств среднего арифметического, связанных с допустимыми преобразованиями шкал. Большинство соотношений (но не все!) между средними арифметическими, используемых в реальных социологических исследованиях, остаются инвариантными относительно положительных линейных преобразований исходных данных – допустимых преобразований интервальных шкал. Таковы, например, соотношения вида:

$\bar{x}_1 \leq \bar{x}_2$

где $x_1$ и $x_2$, средние арифметические значения рассматриваемого признака, вычисленные для каких-либо двух подсовокупностей объектов (подробнее об этом см., например, [Клигер и др., 1978; Орлов, 1985]). Другими словами, большинство соотношений, включающих в себя среднее арифметическое, являются формально адекватными для интервальных шкал. Нетрудно показать, что для порядковых шкал, напротив, большинство подобных соотношений не будут формально адекватными (см. там же). Казалось бы, очевидным является и такое же утверждение для номинальных шкал. Но здесь требуется оговорить один момент.

Конечно, использование среднего арифметического, скажем, для чисел – кодов профессий респондента является бессмысленным. Тем не менее, бывают случаи, когда и для номинальных данных оказывается возможным использование этой статистики. Мы имеем в виду дихотомические номинальные признаки, принимающие два значения – 0 и 1. В соответствующей ситуации становится реальной вполне разумная интерпретация рассматриваемой статистики. Поясним это на примере.

Рассмотрим самый популярный дихотомический признак – пол респондента: 0 - мужчина, 1 - женщина.

Предположим, что у нас 10 респондентов со следующими значениями пола:

0, 0, 1, 1, 1, 0, 0, 0, 0, 1.

Нетрудно видеть, что соответствующее среднее арифметическое равно 0,4. Если мы будем его интерпретировать так, как обычно интерпретируют эту статистику, т.е. как пол некоего “среднего человека”, наиболее типичного представителя совокупности, то мы вряд ли получим что-либо осмысленное: наиболее типичным представителем совокупности, состоящей из здоровых мужчин и женщин, является человек, на 40% являющийся женщиной, на 60% мужчиной? Но оказывается, что возможна еще одна довольно естественная интерпретация нашего значения среднего арифметического: оно означает, что в изучаемой совокупности имеется 40% людей с единичным значением рассматриваемого признака (в данном случае - 40% женщин). Такой интерпретацией вполне можно пользоваться, не рискуя придти к нелепости.

Описанная ситуация весьма существенна для социолога. Как мы покажем ниже (см. п. 2.5 раздела 2), не только средние арифметические, но и многие другие статистики, вычисленные для дихотомических данных, поддаются столь же естественной интерпретации в виде некоторых процентов. А это дает основания использовать “числовой” анализ данных для изучения номинальной информации.

Как известно, формальной адекватности метода недостаточно для того, чтобы его можно было считать подходящим для решения той или иной конкретной задачи. Помимо формальной, требуется еще и *содержательная адекватность.* Метод, подходящий для используемых шкал, может не быть пригодным из содержательных соображений. Это касается и столь простых методов, как методы расчета мер средней тенденции. Пример был приведен в п.5.1 части I.

Содержательное сравнение описанных мер средней тенденции осуществляется во многих работах (см., например [Рабочая книга ..., 1983; Гласс и Стэнли, 1976 ].

Наконец, последнее методическое положение, которое мы упомянем – это необходимость анализа *модели, заложенной в методе.* Применительно к мерам средней тенденции такие модели фактически уже были рассмотрены: эти модели включали в себя предположения о типе шкалы, отвечающей рассматриваемому признаку, о непрерывности признака, о расположении его значений внутри каждого интервала и т.д.

## 1.3. Меры разброса и отвечающие им модели

### 1.3.1. Необходимость введения мер разброса

Прежде всего отметим, что, используя для описания выборки только ту или иную меру средней тенденции, исследователь рискует сильно ошибиться в своей оценке характера изучаемой совокупности респондентов. Например, если изучаемый признак – возраст, то две совокупности людей из 6-ти человек каждая, характеризующиеся следующими значениями возраста, будут иметь одинаковое среднее арифметическое:

10, 10, 10, 50, 50, 50

30, 30, 30, 30, 30, 30.

В то же время совершенно ясно, что практически для любой социологической задачи это будут совсем разные совокупности. И узнать это можно, только как-то оценив степень разброса значений возраста в каждой из них: в первой – разброс большой, во второй – он отсутствует.

Способов оценки степени разброса существует много. Выбор их в первую очередь зависит от типа используемых шкал.

### 1.3.2 Дисперсия. Квантильные размахи

Из математической статистики известно, что самой известной мерой разброса количественного признака является его дисперсия:

$$\sigma^2 = \frac{((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_N - \overline{x})^2)}{N - 1}$$

(напомним, что в знаменателе величина объема выборки уменьшается на единицу для того, чтобы сделать соответствующую точечную выборочную оценку дисперсии несмещенной; σ – среднее квадратическое отклонение). Ясно, что эта статистика может быть формально адекватной только для интервальных шкал (хотя бы потому, что только при этом условии разумно использование среднего арифметического).

Для порядковых шкал обычно используют какие-либо разницы между квантилями. Например, употребительной мерой является квартильный размах: $Q_3$ - $Q_1$. Но, строго говоря, это некорректно, поскольку для порядковой шкалы разности между шкальными значениями не являются осмысленными.

Представляется, что прежде, чем переходить к описанию мер разброса для номинальных признаков, необходимо пояснить, каков “физический” смысл таких мер.

### 1.3.3. Интуитивное представление о разбросе значений номинального признака.

Ясно, что для номинальных признаков некорректным является использование всех приведенных выше мер разброса. Попытаемся понять, как можно интерпретировать такой разброс. Предположим, что в аудитории сидят 100 человек, на которых могут быть надеты свитеры пяти разных расцветок: синие, красные, белые, желтые и зеленые. Вероятно, естественно предполагать, что разброс значений признака “цвет свитера человека” минимален (отсутствует), когда все люди одеты в свитеры одного цвета. Максимальным же разброс естественно считать в том случае, когда все цвета встречаются одинаково часто: 20 человек одеты в синие свитера, 20 человек – в красные и т.д. Другими словами максимальным разброс целесообразно считать при равномерном распределении. Чем ближе распределение к равномерному – тем разброс больше, чем дольше от равномерного – тем разброс меньше.

Известны по крайней мере две меры разброса, опирающиеся на этот принцип – мера качественной вариации и энтропийный коэффициент разброса.

### 1.3.4. Мера качественной вариации.

Чтобы прояснить смысл рассматриваемой меры, прибегнем к упрощенному примеру с дихотомическим признаком. Предположим, что мы организовали танцевальный кружок из 10 человек и пытаемся путем перебора различных вариантов формирования разнополых пар найти такие, в которых мужчина и женщина наиболее удачно подходят друг другу как танцоры. Рассмотрим варианты, отраженные в таблице 3.

Мы видим, что наибольшее количество пар можно организовать, когда распределение по полу равномерно (т.е. количество мужчин равно количеству женщин) или, в соответствии с приведенными выше рассуждениями, когда разброс членов кружка по полу максимален. Более внимательное рассмотрение таблицы

*Таблица 3*

**Зависимость количества пар из разнородных элементов от степени однородности распределения**

| Количество мужчин в кружке | Количество женщин в кружке | Количество возможных танцевальных пар |
|---|---|---|
| 0 | 10 | 0 |
| 1 | 9 | 9 |
| 2 | 8 | 16 |
| 3 | 7 | 21 |
| 4 | 6 | 24 |
| 5 | 5 | 25 |
| 6 | 4 | 24 |
| 7 | 3 | 21 |
| 8 | 2 | 16 |
| 9 | 1 | 9 |
| 10 | 0 | 0 |

позволяет придти к выводу о том, что уровень разброса респондентов по полу и в остальных случаях четко коррелирует с количеством пар из разнородных элементов: чем больше разброс,

тем больше пар можно составить. Рассматриваемая мера разброса – мера качественной вариации – опирается именно на это обстоятельство: ее “ядро” составляет величина, равная количеству упомянутых пар. Поясним на примере способ расчета этой меры (табл.4).

*Таблица 4*

**Частотная таблица для расчета коэффициента качественной вариации**

| Наименование градации рассматриваемого номинального признака | A | B | C |
|---|---|---|---|
| Частота встречаемости градации | 30 | 20 | 70 |

Вычислим коэффициент по следующей формуле:

$$J = \frac{(30 \times 20 + 30 \times 70 + 20 \times 70}{(40 \times 40 + 40 \times 40 + 40 \times 40)}$$

Нетрудно видеть, что в числителе дроби стоит число, равное количеству пар, которые можно составить из разнокачественных элементов: произведение 30×20 – количество пар, первый элемент который обладает свойством А, а второй – свойством В; 30×70 – то же для свойств А и С; 20×70 – для свойств В и С. Другими словами, числитель отражает существо нашего понимания разброса.

Однако считать, что числитель может служить мерой разброса - нельзя. Границы его изменения зависят от объема выборки, от величины конкретных частот. Поэтому, ограничившись числителем, мы тем самым потеряли бы возможность сравнивать меры разброса для разных совокупностей: число, отвечающее большому разбросу в малой выборке, вполне может говорить о весьма несущественном разбросе в большой выборке. Это недопустимо, поскольку, как мы уже отмечали, любой анализ данных связан прежде всего со сравнением разных совокупностей объектов.

Покажем на примере, что максимальное значение числителя рассматриваемой дроби действительно зависит от величин конкретных используемых частот и поэтому числитель не может использоваться в качестве меры разброса. Рассмотрим две частотные таблицы - ту же, которую рассматривали выше и другую, отличающуюся от первой уменьшением всех частот в 10 раз. Другими словами, рассмотрим две разные выборки, характеристики которых отражены в таблице 5.

*Таблица 5*

**Данные, иллюстрирующие зависимость величины меры качественной вариации от объема выборки**

| Наименование градации рассматриваемого признака | Число респондентов (частота) в первой выборке (120 человек) | Гипотетические частоты, отвечающие максимальному значению J | Число респондентов (частота) во второй выборке (12 человек) | Гипотетические частоты, отвечающие максимальному значению J |
|---|---|---|---|---|
| A | 30 | 40 | 3 | 4 |
| B | 20 | 40 | 2 | 4 |
| C | 70 | 40 | 7 | 4 |

При объеме выборки в 12 человек (и, конечно, при трех градациях признака) максимальное количество пар из разнородных элементов равно (4×4 + 4×4 + 4×4) = 48. И реализация такой возможности (отвечающая последнему столбцу таблицы) говорит о наличии максимального разброса по рассматриваемому признаку. Другими словами, для выборки в 12 человек число 48 говорит о максимальном разбросе. А при объеме выборки в 120 человек (при тех же трех градациях) такого малого количества пар не может быть даже при самом минимальном (но ненулевом) разбросе. Ясно, такой минимальный разброс будет иметь место, если какое-то одно значение будет встречаться 119 раз, а другое – один раз (при отсутствии третьего значения). Количество же пар из разнородных элементов в таком случае будет равно 119, что больше 48.

Итак, если мы будем пользоваться только числителем дроби, выражающей коэффициент J, то в одном случае число 48 будет говорить о максимальном разбросе, а в другом – число 119 – о практическом отсутствии разброса. Мы полностью теряем возможность сравнивать величину коэффициента для разных совокупностей. Это вряд ли может быть приемлемо: любой анализ – это сравнение.

Именно для того, чтобы избежать описанного недоразумения, обычно поступают таким образом: в числитель помещают формулу, выражающую суть строящегося коэффициента, а в знаменатель – максимально возможное значение этого коэффициента для рассматриваемой ситуации (в нашем случае эта ситуация определяется объемом выборки и количеством градаций рассматриваемого признака). В итоге получившийся показатель “загоняется” в интервал от 0 до 1 (иногда используется интервал от -1 до +1, как в случае многих коэффициентов связи, начиная с известного коэффициента корреляции). Такая процедура называется нормировкой коэффициента.

Нетрудно проверить, что в рассматриваемом случае описанная нормировка есть деление числителя на аналогичную сумму произведений, отвечающую равномерному распределению (т.е. распределению, когда все градации признака встречаются с одинаковой частотой). Именно это отвечает приведенной выше формуле для вычисления J.

Строгое доказательство того, что именно в случае равномерного распределения число возможных пар рассматриваемого вида будет максимальным, можно найти в [Паниотто, Максименко, 1982]; там же приведена общая формула для коэффициента J (в названной работе он обозначен символом $\alpha_k$):

$$J = \frac{2K}{N^2(k-1)} \cdot \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j$$

где N - объем выборки, k - количество градаций рассматриваемого признака, $n_i$ и $n_j$ - соответственно, частоты встречаемости i -й и j -й градаций.

В заключение обсуждения вопроса о коэффициенте качественной вариации отметим следующий важный для дальнейшего факт. Если мы имеем дело с дихотомическим признаком, принимающим два значения – 0 и 1, то, вычислив для такого признака обычную дисперсию, мы фактически получим соответствующий коэффициент качественной вариации (точнее, величину, равную этому коэффициенту, деленному на 4; предлагаем читателю самому это проверить). Этот факт подтверждает то, что далее станет для нас очень важным: для анализа дихотомических номинальных данных оказывается возможным использование “количественных” методов.

Еще один коэффициент разброса, также подходящий для анализа номинальных данных, основан на понятии энтропии распределения, к рассмотрению которой мы переходим.

### 1.3.5. Определение энтропии. Ее “социологический” смысл. Энтропийный коэффициент разброса

Понятие энтропии всем знакомо по философской, физической, научно-популярной, научно-фантастической литературе – рост энтропии приводит к тепловой смерти вселенной (напомним, что это утверждение связано с идеями статистической термодинамики) и т.д. Мы коснемся этого понятия в очень слабой степени, рассмотрев, как с его помощью характеризуется упомянутая мера неопределенности.

Известно, что степень неопределенности распределения некоторой случайной величины Y (точнее, меры той неопределенности, которую имеет исследователь в смысле знания значения Y для какого-либо случайно выбранного объекта), определяется с помощью энтропии этого распределения. Введем соответствующие определения.

Пусть случайная величина Y принимает конечное число значений 1,2, ..., k с вероятностями, равными, соответственно, $P_1$, $P_2$, ..., $P_k$. (Напомним, что вероятность какого-либо значения для выборки отождествляется с относительной частотой встречаемости этого значения). Введем обозначение:

$P_j = P\ (Y = j)$

Энтропией случайной величины Y (или соответствующего распределения; напомним, что случайная величина отождествляется с отвечающими ей распределением вероятностей) Y называется функция

$$H(Y) = -\sum_{j=1}^{K} P_j \log P_j$$ (основание логарифма произвольно)

(Последняя формула обычно называется формулой Больцмана (Людвиг Больцман, 1844 - 1906 – австрийский физик, основатель статистической термодинамики). Именно формула, связывающая энтропию с термодинамической вероятностью, выгравирована на памятнике Больцману в Вене. Это соотношение дает статистическое обоснование второму началу термодинамики и является основой статистической физики.)

Чтобы лучше раскрыть смысл энтропии, представляется целесообразным пояснить, какого рода содержательные соображения о понятии неопределенности распределения могут навести на мысль об измерении этого понятия с помощью логарифма. Используем рассуждение из [Яглом, Яглом, 1969.С. 45].

Пусть некие независимые друг от друга признаки U и V принимают, соответственно, k и l равновероятностных значений. Рассмотрим, каким свойствам должна удовлетворять некая функция f, характеризующая неопределенность распределений рассматриваемых признаков. Ясно, что f = f (k) (т.е. рассматриваемая функция зависит от числа градаций того признака, неопределенность распределения которого она измеряет) и что f (1) = 0. Очевидно также, что при k >l должно быть справедливо неравенство f (k) > f (l). Число сочетаний значений рассматриваемых признаков равно произведению kl. Естественно полагать, что степень неопределенности двумерного распределения, f (kl) должна быть равна сумме неопределенностей соответствующих одномерных распределений, т.е. f (k l) = f (k) + f (l).

Можно показать, что логарифмическая функция является единственной функцией аргумента k, удовлетворяющей условиям: f (k l) = f (k) + f (l), f (1) =0, f(k) >f (l) при k >l .)

Функция H (Y) и служит мерой неопределенности распределения Y.

(представляется очевидным, почему основание логарифма произвольно; как известно из школьной математики, от одного основания можно легко перейти к другому; все интересующие нас формулы при этом будут отличаться только на некоторый постоянный множитель, что несущественно для их интерпретации).

Чтобы лучше понять смысл энтропии, вникнем в смысл двух следующих ее свойств.

1) H (Y) ≥ 0. Равенство достигается тогда, когда Y принимает только одно значение. Это – ситуация максимальной определенности: случайным образом выбрав объект, мы точно можем сказать, что для него рассматриваемый признак принимает упомянутое значение. Распределение Y выглядит следующим образом:

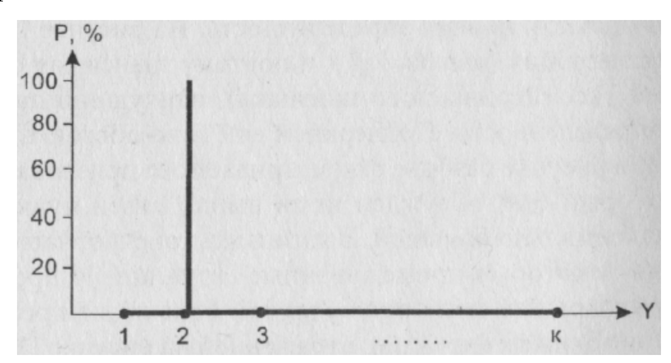



**Рис. 12. Пример распределения с нулевой энтропией**

Единственная отличная от нуля вероятность здесь равна 1. Нетрудно проверить, что для такого распределения энтропия действительно равна нулю.

2) При фиксированном “k” значение энтропии максимально, когда все возможные значения Y равновероятны. Это – ситуация максимальной неопределенности. Предположим, например, что k=5. Тогда распределение Y для такой ситуации будет выглядеть следующим образом:

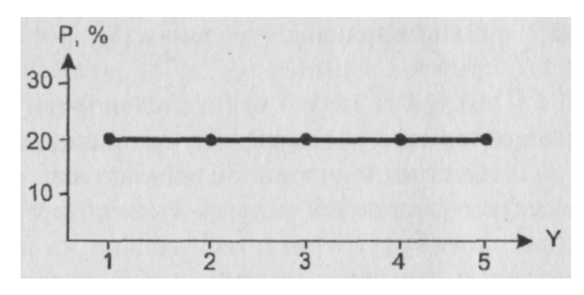


**Рис. 13. Пример распределения с максимальной энтропией при заданном числе градаций признака**

Ясно, что здесь $P_j = 0{,}2$. Нетрудно проверить, что значение энтропии при этом равно log 5, а в общем случае в ситуации полной неопределенности энтропия равна log k. Таким образом, чем больше градаций имеет рассматриваемый признак, тем в принципе большей энтропии может достичь отвечающее ему распределение.

Итак, на рис. 12 – минимальная (нулевая) энтропия, наилучший прогноз, полная определенность. На рис.13 – максимальная энтропия (равная log k и поэтому зависящая от числа градаций рассматриваемого признака), наихудший прогноз, полная неопределенность. Подчеркнем еще и то обстоятельство, что на первом рисунке разброс рассматриваемого признака (в том смысле, который был обсужден нами выше) равен нулю, а на втором – максимально большой. В жизни же, конечно, чаще всего встречаются некоторые промежуточные ситуации. И представляется очевидным, что энтропия будет тем больше, чем реальное распределение ближе к ситуации, отраженной на рис. 13, и тем меньше, чем оно ближе к ситуации, отраженной на рис. 12.

Поэтому будем считать интуитивно ясным тот факт, что энтропия может использоваться при оценке степени разброса значений номинального признака. Однако мы уже упоминали, что максимальное значение энтропии для распределения какого-либо признака зависит от числа его градаций. Следуя той же логике, что была использована нами выше, нетрудно придти к выводу, что сама энтропия, в силу сказанного, не может выступать в качестве меры разброса. Чтобы такое использование было правомерным, значение энтропии необходимо нормировать – поделить на величину максимальной энтропии. Так обычно и поступают: в качестве меры разброса используют энтропийный коэффициент

$$\varepsilon = \frac{H}{H_{\max}} = \frac{H}{\log k}$$

Подробнее об этом см. работу [Паниотто, Максименко, 1982].

В заключение параграфа отметим, что в том направлении науки, которое связано с моделированем социальных процессов, понятие энтропии занимает существенное место. Причины этого нетрудно понять. Скажем, известно, что общества слишком однородные, либо слишком разнородные не является устойчивымы. А однородность может оцениваться как раз с помощью энтропии. Правда, для того, чтобы энтропия могла “работать на прогноз”, необходимо

решить серьезные содержательные вопросы и, в первую очередь, определить – для каких признаков энтропию надо измерять.

# 2. АНАЛИЗ СВЯЗЕЙ МЕЖДУ НОМИНАЛЬНЫМИ ПРИЗНАКАМИ

## 2.1. Анализ номинальных данных как одна из главных задач социолога

В данном параграфе мы коротко покажем, что номинальные данные - главный интересующий социолога вид исходной информации; а анализ связей между признаками - главный вид задач, встречающийся практически в любом эмпирическом социологическом исследовании.

### 2.1.1. Роль номинальных данных в социологии

Роль номинальных данных в социологии огромна. Объяснить это можно следующими (взаимосвязанными) причинами.

Во-первых, именно номинальные данные чаще всего используются социологами. Вероятно, это можно объяснить сравнительной простотой их получения, естественностью интерпретации, интуитивной уверенностью в состоятельности последней.

Во-вторых, номинальные данные являются более надёжными, чем данные, полученные по шкалам более высокого типа, в том смысле, что за ними обычно не стоят трудно проверяемые модели восприятия (имеется в виду восприятие респондентом предлагаемых ему для оценки объектов, суждений, мнений и т.д.; о моделях, предполагаемых известными методами шкалирования, см., например, [Толстова, 1998]), и, в соответствии с этим, при их интерпретации не используются сложные и зачастую сомнительные допущения.

164

В-третьих, в методах, используемых для анализа номинальных данных, обычно бывают "заложены" модели, не вызывающие сомнения, отвечающие естественной логике социолога, изучающего собранную информацию "вручную", без использования математики и ЭВМ. Надеемся, что все сказанное ниже позволит читателю в этом убедиться.

Здесь сделаем небольшое отступление. Среди социологов бытует мнение о том, что достижение интервального уровня измерения всегда является желаемым, поскольку расширяет возможности исследователя, давая ему основания использовать традиционные методы математико-статистического анализа данных. С одной стороны, это, конечно, так: подобные основания действительно имеют под собой почву (хотя надо иметь в виду, что и интервальные

данные - не совсем числовые и поэтому к ним применимы не все упомянутые традиционные алгоритмы). Но, с другой стороны, остается вопрос о том, не слишком ли дорога соответствующая цена, не обесценивается ли полученное преимущество несостоятельностью анализируемых данных. Последнее соображение настолько важно, что некоторые авторы вообще полагают, что в социологии только номинальные шкалы имеют право на существование [Чесноков, 1986]. И принять это соображение во внимание имеет смысл еще и потому, что для анализа номинальных данных имеется много достаточно эффективных методов.

### 2.1.2. Соотношение между причинно-следственными отношениями и формальными методами их изучения

Изучение связей между переменными, как правило, интересует исследователя не само по себе, а как отражение соответствующих причинно-следственных отношений. Представляется излишним доказательство актуальности соответствующих задач, их важность для любого социологического исследования. Однако причинные отношения при изучении социальных явлений не удается выделить в “чистом” виде. Социолог может наблюдать только соответствующие статистические закономерности (статистические

165

связи), в качестве измерителей которых и выступают известные показатели связи (далее мы увидим, в чем именно проявляется статистичность интересующих нас связей). То устойчивое, необходимое, что скрывает за каждым коэффициентом (или за системой таких коэффициентров) зачастую оказывается возможным отождествить с соответствующей причинной зависимостью.

Подчеркнем, однако, понятия "причина" и "следствие" в принципе не могут быть формализованы. Никакая математика не может нам доказать, что такой-то признак служит причиной (следствием) того или иного явления. Можно привести массу примеров, когда наличие даже самой сильной статистической связи совершенно не означает наличие соответствующей причинной зависимости. Например, у людей, как правило, одновременно появляется желание надеть легкое платье и пойти искупаться не потому, что одно причинно обусловливает другое, а потому, что оба эти желания вызваны одним и тем же обстоятельством – наступлением жаркой погоды. Другой пример: два студента одновременно вдруг проявляют необыкновенную тягу к знаниям или, напротив, стремятся отлынивать от занятий не потому, что один на другого причинно воздействует, а потому, что сессия у них в одно и то же время – одновременное причинное воздействие третьего признака на каждый из двух данных вызывает

статистическую связь между данными признаками. Подобные статистические, не являющиеся причинно-следственными, связи в литературе носят название ложной корреляции. Название не очень удачное – корреляция-то (т.е. статистическая связь) как раз истинна, ложно – причинно-следственное отношение.

Итак, математические методы могут лишь навести нас на мысль о существовании причинных отношений, заставить быть более уверенными в своих предположениях, или, напротив, усомниться в них, скорректировать свои априорные представления или даже совсем отказаться от них. Тем не менее, термины "причина" и "следствие" часто употребляются при математическом анализе социологических данных. Однако обычно они отражают лишь априорные исследовательские предположения соответствующего плана.

166

Правда, в одной из известных ветвей многомерного статистического анализа – т.н. причинном (путевом) анализе [Хейс, 1981] термин "причина" используется именно как нечто формально недоказуемое. В его рамках специально изучаются ситуации с ложными корреляциями, подробно рассматривается, как сложные, опосредованные цепочки причинных отношений могут объяснять их наличие, позволяет понять, за счет чего иногда между какими-то признаками может быть сильная статистическая зависимость при полном отсутствии причинно-следственной, какими сложными опосредованными причинными отношениями эта связь может объясняться.

### 2.1.3. О понятии таблицы сопряженности.

Представляется естественным использовать для оценки связей между признаками т. н. частотные таблицы, или таблицы сопряженности (по существу мы о них уже говорили – это выборочные оценки вероятностных распределений многомерных случайных величин; так, в таблице 3 части I приведен пример распределения для двумерной величины). Заметим, что последний термин обязан своим происхождением именно тому обстоятельству, что на основе анализа подобных таблиц можно судить о сопряженности (совместной встречаемости) каких-то значений одних признаков с некоторыми значениями других признаков. Как мы увидим, связь между номинальными признаками, собственно говоря, и выражается в виде подобных сопряженностей.

Предположим, что мы имеем два признака X и Y, первый из которых принимает "r" значений 1, 2, . . ., r, а второй – "c" значений 1, 2, . . ., c. Назовем двумерной таблицей сопряженности (двумерной частотной таблицей) некоторую матрицу, на пересечении i-й строки

и j-го столбца которой стоит число $n_{iij}$, означающее количество объектов, обладающих i-м значением первого признака и j-м значением второго (i =1, . . ., r; j =1, . . ., c) (использование латинских букв r и c в указанном смысле принято в литературе; эти буквы сопрягаются с английским словами raw и column, означающими "строка" и "столбец" соответственно; это не позволяет нам забывать, что значения одного признака отвечают строкам таблицы сопряженности, а другого - столбцам). Другими словами, таблица сопряженности выглядит так:

*Таблица 6.*

**Общий вид таблицы сопряженности**

$$\|n_{ij}\| = \begin{Vmatrix} n_{11} & n_{12} & \dots & n_{1c} \\ n_{21} & n_{22} & \dots & n_{2c} \\ \dots & \dots & \dots & \dots \\ n_{r1} & n_{r2} & \dots & n_{rc} \end{Vmatrix}$$

Обычно ее представляют в несколько ином виде, с явно обозначенными наименованиями признаков и их значений и выписанными маргинальными суммами:

*Таблица 7*

**Общий вид таблицы сопряженности**

| X | Y | | | | | | Маргиналы по строкам |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | … | j | … | c | |
| 1 | $n_{11}$ | $n_{12}$ | … | $n_{1j}$ | … | $n_{1c}$ | $n_{1.}$ |
| 2 | $n_{21}$ | $n_{22}$ | … | $n_{2j}$ | … | $n_{2c}$ | $n_{2.}$ |
| … | … | … | … | … | … | … | … |
| i | $n_{i1}$ | $n_{i2}$ | … | $n_{ij}$ | … | $n_{ic}$ | $n_{i.}$ |
| … | … | … | … | … | … | … | … |
| r | $n_{r1}$ | $n_{r2}$ | … | $n_{rj}$ | … | $n_{rc}$ | $n_{r.}$ |
| Маргиналы по столбцам | $n_{.1}$ | $n_{.2}$ | … | $n_{.j}$ | … | $n_{.c}$ | n |

Правый крайний столбец образуют строковые маргинальные суммы (маргиналы по строкам). Величина $n_{i.}$ равна сумме элементов i-й строки (т.е. числу тех объектов, для которых первый признак принимает значение i). Нижняя строка образуется столбцовыми маргинальными суммами (маргиналами по столбцам). Величина $n_{.j}$ равна сумме элементов j-го столбца (т.е. числу тех объектов, для которых второй признак принимает значение j). n - объем выборки, он равен сумме маргиналов по столбцам (либо по строкам).

В последние годы в литературе все более используется расширительное понимание таблицы сопряженности. Предполагается, что в качестве ее элементов могут фигурировать не только частоты, но и многие другие числа: скажем, в клетках половозрастной таблицы могут стоять средние значения зарплаты тех людей, которые характеризуются отвечающим клетке значениям пола и возраста. Таким же образом в клетки таблицы могут быть помещены средние другого рода (мода, медиана), дисперсии, величины отклонений от средних по строке (столбцу), разница между эмпирической и теоретической частотой (см. п.2.2.1) и т.д. (см., например, [Ростовцев и др., 1997. С.177-179]). О том же расширительном понимании таблицы сопряженности говорится в описании известного пакета SPSS.

Ниже, приводя примеры, под объектами, число которых подсчитывается при построении таблицы сопряженности, мы будем иметь в виду респондентов. Хотелось бы, чтобы читатель давал себе отчет в условности таких примеров, понимая, что отнюдь не только респонденты могут интересовать социолога.

## 2.2. Классификация задач анализа связей номинальных признаков

### 2.2.1. Диалектика в понимании признака и его значений.

Со следующей главы мы начнем описание ряда методов анализа номинальных данных. Придадим цельность нашему изложению путем установления связи между этими методами посредством прослеживания определенного родства заложенных в этих методах моделей. Сделаем это посредством выработки единого основания для классификации всех рассматриваемых алгоритмов, основания, связанного с определенной типологией социологических задач.

Предлагаемое основание будет опираться на то обстоятельство, что для социолога важно осознание необходимости определенной

169

диалектики в понимании признака и его значений: выделение ситуаций, когда отдельной альтернативе имеет смысл придать статус самостоятельного признака.

Приведем пример. Нас может интересовать, каким является отвечающее респонденту значение признака "профессия", а может – является ли этот респондент или не является учителем. Во втором случае мы придали статус признака одному значению признака

"профессия" – тому, которое называлось "учитель". К такому переходу нас подталкивает не желание пооригинальничать, а стремление адекватно решать стоящие перед социологом задачи. Скажем, изучая связи между рассматриваемыми переменными, мы можем придти к выводу, что профессия никак не связана с полом (забегая вперед, скажем, что такой вывод можно сделать, использовав какой-либо из известных коэффициентов связи, рассчитывающихся на базе таблицы сопряженности "пол – профессия", скажем, критерий "Хи-квадрат", см. п. 2.3.1). Тем не менее, та же статистика может нам говорить, что почти все учителя – женщины, т.е. что соответствующее отдельное значение признака "профессия" связано с полом. Чтобы не "упустить" эту "локальную" связь, мы и должны рассмотреть отдельный дихотомический признак "быть учителем" с целью измерения величины его связи с признаком "пол".

Описанное требование можно обобщить: самостоятельной переменной может отвечать не одно значение некоторого признака, а сочетание таких значений (скажем, при решении ряда задач имеет смысл объединить, учителей и врачей вместе), каждое из которых соответствует, вообще говоря, своему признаку (о таких ситуациях, когда объединяются альтернативы разных признаков, пойдет речь в п.2.5).

Два слова о терминах. В работе [Чесноков, 1982] предлагается называть *глобальными* коэффициенты парной связи, рассчитывающиеся на основе учета всех градаций рассматриваемых признаков, и *локальными* – коэффициенты связи, рассчитывающиеся на основе учета одной градации одного признака и одной градации другого. Нам представляется неприемлемым деление всех показателей на глобальные и локальные, поскольку при таком

170

подходе из рассмотрения (во всяком случае на терминологическом уровне), выпадают связи "промежуточных" видов: такие, когда учитываются несколько градаций каждого признака. Однако термин “локальная связь” мы будем использовать, понимая под таковой связь между отдельными альтернативами.

Заметим, что приведенные выше соображения имеют самое непосредственное отношение к проблеме социологического измерения, к анализу понятия "признак" и, в конечном счете, к проблеме операционализации понятий, к изучению перехода от реальных многогранных объектов к их узкому, всегда ограниченному описанию набором некоторых признаков (к "мышлению признаками", по выражению автора работы [Ноэль, 1993]).

Описанные ситуации возникают в силу того, что, с одной стороны, само понятие признака имеет смысл только при некоторой однокачественности тех объектов, для которых значения признаков вычисляются; с другой стороны, – каждому значению признака отвечает

свое собственное качество. Понятие однокачественности относительно. На разных этапах исследования может возникнуть потребность однокачественные объекты считать разнокачественными и наоборот. Так, выше мы показали, что бывают ситуации, когда однокачественными объектами мы считаем всех тех и только тех респондентов, которые имеют профессию учителя. Человек же с профессией врача в такой ситуации будет иметь другое качество. При изучении проблем интеллигенции учитель и врач могут стать однокачественными объектами. Если же мы работаем с признаком "профессия" как единым целым, то тем самым полагаем, что этот признак отражает существование некоторого социального института и однокачественными являются все члены такого общества, в котором этот институт имеется.

В обосновании необходимости "склеивания" отдельных значений разных (вообще говоря) признаков просматривается актуальность решения следующей проблемы социологического измерения: чтобы отразить латентные свойства объекта, мы вынуждены "выдергивать" отдельные значения разных признаков, формировать из этих "надерганных" значений различные

171

комбинации, надеясь, что какое-то сочетание хотя бы частично явится индикатором определенного "поведения" объекта.

Дальнейшее обобщение требования склеивания отдельных градаций приводит к осознанию возможности рассмотрения в качестве нового признака не сочетания отдельных альтернатив, а сочетания нескольких признаков. Соответствующее обобщение проблемы измерения очевидно: новым измеряемым признаком является здесь комбинация исходных признаков.

Продолжая ту же логику, естественно приходим к необходимости рассмотрения всех признаков сразу как единой системы.

Выделение перечисленных возможностей мы будем рассматривать как основу для дальнейшего изложения (в частности, для классификации методов анализа связей номинальных признаков).

Итак, в соответствии с предлагаемой точкой зрения, каждый рассматриваемый метод можно трактовать как реализацию следующего процесса: все исходные номинальные признаки как бы "рассыпаются" на отдельные градации, которые затем по-разному комбинируются, на их основе строятся новые признаки, взаимоотношения которых далее изучаются. Каждый метод анализа связей номинальных данных предлагается рассматривать как метод поиска либо связей между разными группами альтернатив, либо групп альтернатив, определяющих некоторое

поведение респондентов (задаваемое разными способами). Методы систематизируются в зависимости от отвечающих им способов агрегирования отдельных альтернатив в новые признаки.

Использование предлагаемого подхода, на наш взгляд, побуждает исследователя не забывать о существовании многих методов, весьма адекватных социологическим задачам, но мало используемых социологами.

В данном разделе мы будем рассматривать методы, которые включаются в указанную классификацию. Но прежде, чем более подробно ее описать (что будет сделано в п. 2.2.2), представляется важным рассмотреть один момент, позволяющий лучше понять, как модели, заложенные в интересующих нас методах, соотносятся с моделями других известных методов анализа данных (о других

172

моментах такого рода см. п. 2.2.3).

Нетрудно заметить, что упомянутые выше задачи (и отвечающие им методы), связанные с поиском групп альтернатив, определяющих некоторое поведение респондентов, очень похожи на задачи поиска того, что в математической статистике (в частности, в дисперсионном и регрессионном анализе; описание первого можно найти, например, в [Статистические методы..., 1979], о втором пойдет речь в п.2.6), называется *взаимодействием*.

Напомним, что использование этого термина предполагает выделение среди всех признаков главного признака (зависимого, выходного, целевого, объясняемого, результирующего, признака-функции, признака-следствия) и группы детерминирующих его признаков (независимых, входных, объясняющих, предикторов, признаков - аргументов, признаков-причин; подробнее о подобных терминах см. п. 2.5.3.1). “Взаимодействие” означает сочетание значений независимых признаков, определяющих тот или иной уровень зависимого (заметим, что в дисперсионном анализе зависимый признак предполагается количественным, т.е. таким, значения которого получены по крайней по интервальной шкале; а совокупность независимых признаков фиксируется). Например, при изучении миграционного поведения взаимодействием может служить свойство респондента одновременно быть мужчиной (т.е. обладать, скажем, значением “1” признака 4 - “пол”) и иметь высшее образование (т.е. обладать, например, значением “5” признака 6 - “образование”), если это свойство детерминирует желание обладающего им человека уехать за границу.

Роль поиска взаимодействий в эмпирической социологии вряд ли можно преувеличить. Однако представляется, что потребность практики делает целесообразным расширение этого

понятия. Для того, чтобы пояснить, каким способом это можно сделать, попытаемся вдуматься в смысл того, что значит делать какие-то выводы в терминах рассматриваемых (номинальных) признаков. Вероятно, исходя из здравого смысла, подобные выводы должны иметь вид (мы имеем в виду формальную структуру того статистического утверждения, которое служит социологу основой для дальнейших выводов о причинно-следственных отношениях):

"5-е значение 8-го признака часто встречается с 3-м значением 14-го и 1-м значением 2-го", "из того, что 3-й признак принимает

173

2-е значение одновременно с тем, что 4-й принимает 5-е значение, как правило, следует, что 6-й признак принимает либо 2-е, либо 3-е", "из того, что 3-й признак принимает какое-либо значение, кроме 2-го, следует, что 7-й признак принимает 4-е значение" и т.д. (надеемся, что для понимания сказанного не требуется более конкретно формулировать подобные утверждения: скажем, указывать, что 3-й признак - это возраст, его 5-е значение - указание того, что возраст конкретного респондента заключён в интервале от 35 до 40 лет и т.д.).

(Выражения, подобные сформулированным, являются наиболее естественными для социолога. Они отвечают сути номинальных шкал, тому, что каждое значение признака означает самостоятельное автономное качество объекта. Однако исследователь зачастую стремится по-другому формулировать искомые содержательные выводы, вольно или невольно вписывая их в традиционные рамки классических математико-статистических формулировок: "такие-то два признака имеют сильную статистическую связь", "второй признак линейно зависит от седьмого" и т.д. Можно показать, что такие формулировки тоже могут быть "переведены" на язык наших взаимодействий.)

Анализ подобного рода выражений заставляет следующим образом *обобщить понятие взаимодействия*:

- совокупность признаков-предикторов будем считать "плавающей" (естественно, - в пределах множества признаков, заданных в исследовании; напомним, что в дисперсионном анализе фиксируется небольшое количество признаков-предикторов и рассматриваются все возможные сочетания их значений; среди этих значений и ищутся взаимодействия); в частности, будем полагать, что какое-то сочетание значений одного набора предикторов может определять одно значение признака-функции, а некоторое сочетание значений другого набора предикторов – другое значение функции; например, в добавление к высказанному выше гипотетическому предположению о том, что у мужчин с высшим образованием появляется желание покинуть

Родину, можно добавить еще одно предположение – о том, что женщины, имеющие более двух детей, напротив, выступают против отъезда за границу;

174

- будем полагать, что взаимодействием может быть не только конъюнкция суждений типа “значение такого-то признака равно тому-то” (именно конъюнкцией суждений “человек – мужчина” и “человек имеет высшее образование” является суждение “человек является мужчиной с высшим образованием”), а любые логические функции от таких выражений (предполагаем, что читатель знает определение основных логических функций - конъюнкции, дизъюнкции, импликации, отрицания; используемые здесь и ниже сведения по логике можно почерпнуть, например, из [Бочаров, Маркин, 1994]); например, взаимодействием будем считать суждение "человек является или пенсионером, или женщиной с маленьким ребенком, или не бизнесменом", если люди, обладающие соответствующими свойствами, не желают покидать родные места; (сравним также с упомянутыми выше "2-м значением 3-го признака и 5-м – 4-го, любым значением 3-го, кроме 2-го";) такого рода функции будем называть *объясняющими, или детерминирующими, положениями (выражениями);* их будем описывать так, как это обычно делается в литературе: используя для обозначения входящих в них признаков букву X с индексами ($X_3(2)$ & $X_4(5)$, $\rceil X_3(2)$ и т.д.).
- будем полагать, что наше взаимодействие может определять не только некоторое значение непрерывного признака (как в дисперсионном анализе), но и любую логическую функцию значений произвольных, в том числе дискретных (в частности, номинальных) признаков (ср. упомянутые выше "3-е значение 14-го признака и 1-е – 2-го; 2-е или 3-е значение 4-го признака); каким-либо другим образом задаваемое "поведение" респондента (примеры будут приведены в п.2.5, при обсуждении алгоритмов THAID и CHAID); частоту в таблице сопряжённости (ср. “ 5-е значение 8-го признака часто встречается с 3-м значением 14-го и 1-м значением 2-го”; это мы рассматривать не будем; однако подчеркнем, что речь идет об очень актуальных для социологии задачах, решаемых с помощью логлинейного анализа [Аптон, 1982]); а может и ничего не определять, но тогда естественно требовать просто истинность взаимодействия как логической функции; то, что определяет взаимодействие, будем называть

175

*объясняемыми, или детерминируемыми, положениями*. Их будем описывать обычно, используя для входящих в них признаков букву Y с индексами;

О поиске обобщенных взаимодействий будем говорить как о поиске закономерностей или детерминаций.

Рассмотрим еще одну сторону понимания термина "взаимодействие" - то, каким образом могут быть связаны объясняющее и объясняемое положения. Обратим внимание на некоторые аспекты приведенных выше формулировок типичных социологических утверждений в терминах используемых номинальных признаков. “5-е значение 8-го признака **часто встречается** с 3-м значением 14-го и 1-м значением 2-го”, “из того, что 3-й признак принимает 2-е значение одновременно с тем, что 4-й принимает 5-е значение, **как правило**, следует, что 6-й признак принимает либо 2-е, либо 3-е”. Представляются очевидными причины появления выделенных слов в приведенных выражениях. Мы имеем дело лишь со статистическими закономерностями, являющимися в определенном смысле приближенными. Например, если даже вполне можно считать, что мужчины с высшим образованием имеют склонность эмигрировать, практически всегда из этого правила будут исключения. И всегда встает вопрос о том, каково должно быть количество подобных исключений для того, чтобы мы все-таки считали найденную закономерность закономерностью. К этому вопросу мы не раз будем возвращаться.

Как формализовать выражения "часто встречается", "как правило" и т.д.? Без формализации мы не можем проверять справедливость рассматриваемых суждений. Формализация же – это фрагмент используемой модели. Он разный в разных методах. Так, в неоднократно упомянутом нами дисперсионном анализе речь идет о статистической значимости различий средних значений выходного признака для респондентов, обладающих разными сочетаниями значений предикторов. Как мы увидим ниже, в других интересующих нас алгоритмах задействованы другие критерии (о них пойдет речь ниже, при описании соответствующих алгоритмов). Возможность разных критериев тоже может рассматриваться как элемент обобщенного подхода к пониманию

176

взаимодействия. Обсуждая подобные критерии, будем говорить о формализации *понятия приближенности* искомой закономерности.

При таком понимании взаимодействия можно сказать, что поиск взаимодействий разного рода служит основой большинства рассматриваемых нами методов анализа номинальных данных. В следующем параграфе будут приведены примеры.

## 2.2. Классификация рассматриваемых задач и отвечающих им методов

Ниже в скобках бы будем указывать примеры математических методов, направленных на решение задач выделяемых классов. При первом чтении это можно опустить. Мы называем конкретные методы уже сейчас, до того как они будут описаны (а следующие параграфы будут посвящены такому описанию; сами названия этих параграфов отвечают названиям выделенных ниже классов задач), по двум причинам: во-первых, для того, чтобы читатель, знакомый с упоминаемыми методами, лучше понял нашу классификацию; во-вторых, мы надеемся,что читатель вернется к настоящему параграфу после прочтения всей книги с целью более четко представить себе совокупность тех алгоритмов, из числа которых ему предстоить выбрать инструмент для обнаружения интересующих его закономерностей.

Итак, в соответствии с предлагаемым основанием выделяются задачи типа:

– "альтернатива-альтернатива", т.е. такие, которые позволяют изучать связь между отдельными значениями любых рассматриваемых признаков (примером является детерминационный анализ [Чесноков, 1982]);

– "(группа альтернатив) - (группа альтернатив)" (анализ фрагментов таблиц сопряженности [Интерпретация и анализ ..., гл. 2], алгоритмы типа "пятна" и "полосы" [Ростовцев, 1985. С. 203-214]); эту группу методов можно расширить, условно назвав результат такого расширения методами типа

– " (группа альтернатив) – ("поведение" объектов)", где “поведение" (подчеркнем, - не одного объекта, а целой

177

совокупности, заданной рассматриваемой группой альтернатив; такое "поведение" в определенном смысле есть описание этой совокупности, которое, в свою очередь, можно интерпретировать как характеристику некоторого типа объектов) может пониматься по-разному: как определенный каким-либо образом "средний" уровень заранее заданного результирующего признака (скажем мы можем искать тип людей с низким уровнем зарплаты и тип людей с высоким уровнем зарплаты), как истинность для рассматриваемой совокупности некоторой логической функции от элементарных формул типа P(a)=1, (так называемых логических закономерностей), где буквой P обозначен произвольный признак, а приведенное выражение означает: "значение признака P для объекта a равно 1" и т.д. (методы выявления логических закономерностей [Лбов, 1981], методы поиска детерминирующих сочетаний значений рассматриваемых признаков, в том числе известные на Западе алгоритмы, для обозначения которых используются аббревиатуры, включающие в себя сочетание AID

(automatic interaction detector): THAID [Интерпретация и анализ данных в социологических исследованиях, 1987, с. 136-151; Messenger, Mandell 1972; Morgan, Messenger, 1973]), CHAID [Agresti, 1990; Magidson, 1993; Derrick, Magidson, 1992], AID3 [Sonquist, Morgan, 1973] и т.д. Сравнение THAID и AID3 осуществляется в [Kass, 1980]. Ряд методов описан в [Типология и классификация в социологических исследованиях, 1982, с. 213-231]. Назовем также брошюру [Ливанова Т. Н. 1990], где подробно описан процесс реализации на ЕС ЭВМ алгоритма AID3. Хотя в наше время персональных компьютеров такое описание не является актуальным, тем не менее, на наш взгляд, указанная работа не стала бесполезной для социолога, поскольку в ней помимо правил обращения с ЭВМ серии ЕС подробно раскрывается сущность самого алгоритма).

Частным случаем упомянутых комбинаций явится объединение в одну группу альтернатив, отвечающих одному признаку. В соответствии с этим, выделим класс задач:

– "признак - признак" (традиционные, наиболее знакомые социологу коэффициенты парной связи).

Продолжая рассуждения, отвечающие той же логике, нетрудно придти к выводу, что та же специфика измерительных процедур

178

может вызвать потребность объединять не только "надерганные" из разных признаков альтернативы, но и признаки в целом. в соответствии с этим, в рамках нашей классификации выделим группы методов:

– "признак - (группа признаков)" (регрессионный анализ, многие методы построения индексов);

(Отметим, что при использовании регрессионного анализа зачастую решаются также задачи типа "(группа альтернатив) - ("поведение" объекта)"; это ярко демонстрирует его так называемый номинальный вариант [Аргунова, 1990; Типология и классификация..., 1982; Hardy, 1993], см. также п. 2.6.)

– "(группа признаков) - (группа признаков)" (канонический анализ [Интерпретация и анализ ..., 1987]). Это известный математико-статистический метод. Однако он крайне редко используется социологами, считающими его типично "количественным" методом. В действительности же соответствующий подход является актуальным для анализа именно номинальных данных: он дает возможность осуществлять их оцифровку (т.е. приписать каждому значению номинального признака некоторое число), изучать связи между признаками с т. н. "совместными" альтернативами, эффективно находить веса признаков при формировании

из них индекса. Идеи, заложенные в каноническом анализе используются в таком широко применяющимся в современной западной социологии (в том числе в ставших “модными” в России маркетинговых исследованиях) методе, как корреспонденс-анализ, или анализ соответствий [Clausen, 1998]).

Тип задач, отвечающих рассмотрению всей совокупности признаков как системы, назовем так:

– анализ системы признаков (логлинейный анализ [Аптон, 1982; Елисеева, Рукавишников, 1977; Мирзоев, 1980,1981; Миркин, 1980 ]; причинный анализ [Елисеева, Рукавишников, 1982; Осипов, Андреев, 1977; Хейс, 1981]).

К сожалению, в настоящей работе мы не имеем возможности рассмотреть последние два типа задач.

Конечно, если строго следовать формальной логике, можно заметить, что почти все упомянутые классы методов могут быть

179

сведены к одному – классу "(группа альтернатив)-(группа альтернатив)", поскольку с формальной точки зрения частным случаем группы альтернатив является и отдельная альтернатива; и набор градаций, отвечающих одному признаку; и совокупности значений сразу нескольких признаков. Но с содержательной точки зрения все же мы не можем игнорировать различие между выделенными выше совокупностями альтернатив. В частности, понятие признака – это нечто, отвечающее вполне определенной социальной реальности. За частью альтернатив признака эта реальность не стоит. И, как мы увидим ниже, методы, позволяющие решать задачи выделенных классов, различны, поскольку различны постановки соответствующих содержательных вопросов.

Казалось бы, изложение надо начинать с описания наиболее простых методов – типа “альтернатива – альтернатива”. Однако исторически сложилось так, что сначала были разработаны коэффициенты парной связи между признаками (т.е. наши методы типа “признак – признак”). А все остальные подходы опирались на соответствующие теоретические положения. Мы не хотим претендовать на разработку новых подходов к обоснованию известных коэффициентов. Поэтому начнем как бы с середины нашей схемы – с описания методов измерения связей между двумя номинальными признаками. Однако прежде позволим себе некоторое отступление от основного содержания настоящей книги. Дело в том, что подходами, рассматриваемыми в настоящей работе, отнюдь не ограничивается ни совокупность всех методов анализа номинальных данных вообще, ни совокупность методов анализа связей между

номинальными переменными. Для того, чтобы более четко охарактеризовать круг задач, решение которых становится доступным с помощью подходов, описанных в следующих параграфах, попытаемся очертить то место, которое эти подходы занимают в гораздо более широкой совокупности известных методов анализа номинальных данных. Сделаем это, обратившись к рассуждениям, нетрадиционным для работ по анализу данных.

180

### 2.2.3. Выделение двух основных групп методов анализа номинальных данных. Место рассматриваемых подходов в этой группировке

Специфичность настоящего параграфа состоит в том, что мы попытаемся достичь сформулированной цели с помощью установления связи между идеями математики и теоретической социологии. Говоря подробнее, мы на примере покажем, что математик зачастую ставит перед собой те же вопросы, что и социолог, но специфика ответов у каждого специалиста (понятия "математик" и "социолог" мы здесь интерпретируем как некоторые идеальные типы, как отражение разницы видения мира разными исследователями, разницы, обусловленной различием их природных данных, склада ума, той среды, в которой они формировались как ученые и т.д.) своя.

"Математик" в большей мере умеет вычленить в реальности какие-то поддающиеся формализации, строгому описанию фрагменты. При этом может не только использовать известный математический язык, но и создавать новый (достаточно формализованное, строгое описание каких-то аспектов реальности, по определению, называется математическим). Ясно, что строгость описания реальности сопряжена со сравнительной ограниченностью, бедностью описываемого. "Социолог" дает более расплывчатое описание увиденного. Но расплывчатость эта зачастую обусловливается более широким кругозором, пониманием того, что отнюдь не все важные для социологии аспекты реальности поддаются формализации, по крайней мере, при современном развитии науки (в свете сказанного представляется очевидной причина того, почему Конт в своей известной классификации наук самой простой наукой назвал математику, а самой сложной – социологию).

Два слова о том, почему мы сочли нужным включить в книгу настоящий параграф. Задуматься о глубинных связях социологии и математики автора побудила необходимость решить известную проблему преподавания студентам-социологам дисциплин,

181

связанных с использованием математического аппарата. Как мы уже отмечали, студенты часто отторгают такие дисциплины, полагая, что они являются чужеродными для социолога. "Противоядием" против такого отторжения обычно служит демонстрация студентам многочисленных примеров использования в эмпирической социологии методов анализа данных (либо методов математического моделирования разного рода социальных явлений и процессов). "Хорошие" студенты начинают понимать, что математика необходима им для будущей практической работы с эмпирическими данными. Однако при этом никакой глубинной связи между социологией и математикой не усматривается. Само собой разумеющимися обычно считаются следующие положения.

(1) Да, математика помогает социологу охватить единым взором огромные массивы, коротко выразить суть содержащихся в них статистических закономерностей, взаимосвязей между отдельными явлениями и т.д. (2) Но к получению наиболее интересных для социолога фактов эмпирической социологии, связанных с серьезным анализом причинно-следственных отношений математика имеет слабое отношение, поскольку она использует методы, разработанные в основном для естественных наук и поэтому позволяет улавливать зависимости, хотя и важные для социолога, но не носящие специфически социологического характера. (3) Более того, к поиску закономерностей, касающихся глубокого анализа сознания респондента, математика вообще не имеет отношения. Этот более глубокий анализ связывается обычно с пониманием, а не с объяснением. Соответствующее знание можно получить только с помощью т.н. качественных методов. (4) Тем более математика далека от того, с чем имеет дело т.н. теоретическая социология.

Определенные размышления позволили нам придти к несогласию с положениями (2), (3), (4). На наш взгляд, связь между математикой и социологией гораздо глубже, чем это принято считать. То, что студенты ее не видят, представляется естественным. Изучением такой связи наша наука практически не занималась. Лишь в самые последние годы в работах специалистов

182

по теоретической социологии стали появляться параграфы с названиями: "Программа статистически-вероятностно ориентированной науки об обществе" (о творчестве Кондорсе), "Идея инкорпорирования учения о социальном прогрессе в математическое естествознание" (о творчестве И.Канта) [Давыдов, 1995]. Однако соответствующий контекст наводит на мысль о том, что эти словосочетания отражают скорее некие интуитивные догадки, пожелания на будущее, чем конструктивный подход к изучению общественных закономерностей с помощью математического аппарата. Ниже мы по существу попытаемся внести некоторый элемент

конструктивности в понимание связи идей математики и теоретической социологии.

Перейдем к выделению интересующих нас групп методов.

Во Введении мы уже предложили некоторую группировку (классификацию) методов анализа данных - деление их на методы дескриптивной статистики, анализа связей между признаками, классификации объектов и поиска латентных переменных. Однако эта классификация является довольно грубой, носит весьма относительный характер и в весьма слабой мере опирается на более или менее серьезные (с точки зрения глубинных моментов, мешающих адекватности использования математики в социологии) модельные предпосылки.

Выделим в огромной совокупности методов анализа номинальных данных два мощных направления, стихийно сложившихся в мировой науке. За каждым из них стоит своя методологическая концепция, свой круг решаемых задач. Глубинные методологические предпосылки, лежащие в основании такого выделения, касаются рефлексии социолога по поводу процесса формирования используемых в исследовании понятий, связаны, в частности, с известным многовековым обсуждением вопросов о номинализме и реализме в социологии. Напомним, о чем идет речь.

Начало упомянутых рассмотрений относится к известному спору об "универсалиях" средневековых схоластов (спор об отношении общего к единичному) [Краткий очерк …, 1960. С.111]. "Реалисты" полагали, что "универсалии" (общие роды) существуют реально, независимо от человеческой мысли и речи.

183

"Номиналисты" – что "универсалии" не существуют реально, не зависимо от человека. Они суть только общие имена (например, "человек вообще", как родовая общность, не существует; реально существуют только отдельные люди; "человек" – лишь общее имя, которым называется каждый конкретный человек).

Среди авторов методов анализа данных также можно выделить своеобразных "реалистов" и "номиналистов". И показать это можно, обратившись к анализу выделяемых нами направлений.

Предлагаемая классификация опирается на некоторые фундаментальные модельные предположения о характере используемых номинальных признаков. Имеется в виду возможность различной интерпретации номинальных данных. Речь идет о том, считаем ли мы, что значения каждого номинального признака являются самостоятельными сущностями, отвечающими разным качествам изучаемых объектов (что часто отождествляется с

"превращением" каждого значения в автономный дихотомический признак; о такой дихотомизации пойдет речь в п. 2.6.3), или же полагаем, что за этими значениями (сочетаниями таких значений) стоит некоторая непрерывная (случайная) величина. В последнем случае мы опираемся на предположение о том, что номинальность наблюдаемого признака объясняется нашим неумением точно измерить "стоящую" за признаком переменную (заметим, что здесь мы не касаемся затронутой выше проблемы, связанной с возможностью рассмотрения каждого найденного с помощью некоторых приемов анализа данных сочетания значений каких-либо признаков как значения строящегося одномерного индекса, см. начало п.2.2.1).

Так, можно рассматривать профессию как единое целое, а можно отдельно рассмотреть свойство "Быть учителем", или свойство "Иметь профессию, представителей которой относят к интеллигенции" т.д.

Выделение указанных подходов к интерпретации номинальных данных представляется достаточно принципиальным по крайней мере по двум причинам.

Первую причину можно назвать *гносеологической*. Именно анализируя возможность усматривать за наблюдаемым

184

признакам некоторую скрытую непрерывную переменную, мы попадаем в самую гущу интересующего нас спора между сторонниками социологического реализма и социологического номинализма Если мы полагаем, что отдельные градации какого-либо признака представляют собой самостоятельные сущности, т.е. отказываемся пользоваться предположением о существовании некоторой переменной, стоящей за ними, то тем самым встаем на сторону номинализма. В таком случае мы полагаем, например, что существуют люди-учителя, люди – токари, а вот понятие "профессия человека" – это лишь некоторое введенное для удобства и лишенное всякого онтологического содержания название совокупности людей, рассматриваемых как носителей указанных свойств. В такой ситуации столь же бессодержательной будет фраза: "пол и профессия статистически связаны друг с другом". Но вполне осмыслено высказывание: "почти все учителя – женщины".

Если же мы считаем, что наблюдаемые значения – это лишь разные проявления некоторой объективно существующей непрерывной латентной переменной, т.е. некоторого общего для всех людей (системного) качества, то тем самым переходим на позиции социологического реализма (во всяком случае, относительно рассматриваемых качеств отдельных людей).

Представляется возможным также связать первую интерпретацию с гуманитарным

подходом к измерению, а вторую – с естественно-научным подходом (об этих подходах см. [Чесноков, 1986]; теория гуманитарных измерений принимает как фундаментальный факт способность людей различать образы и поименовывать их).

Таким образом, мы видим, что одна из актуальных для социологии проблем своеобразно, в каком-то узком своем аспекте, рассматривается математикой

Вторая причина выделения названных подходов к интерпретации номинальных данных – чисто *практическая.* Разные интерпретации приводят к возможности постановки разных задач и, соответственно, – к возникновению (и использованию) разных методов анализа данных.

185

Первая интерпретация обусловливает то, что во главу угла исследователь ставит поиск сочетаний значений признаков, детерминирующих "поведение" (по-разному понимаемое) респондента, т.е. поиск взаимодействий. Соответствующим методам мы уделим большое внимание.

При второй интерпретации действия исследователя, как правило, бывают направлены на то, чтобы “вытащить” из исходной информации "стоящую за кадром" латентную переменную, найти “истинное” ее значение для каждого респондента. Часто при этом используются идеи т.н. "оцифровки", т.е. приписывания каждой градации любого номинального (порядкового) признака определенного числа, отвечающего искомому "истинному" значению соответствующей латентной переменной. Речь идет о широком круге родственных друг другу статистических методов, активно применяющихся в западной социологии (особенно во Франции, где совокупность этих методов зачастую отождествляется с методами анализа данных), но слабо известных российским социологам. Это анализ соответствий [Адамов, 1991; Дидэ, 1979, 1985; Жамбю, 1978, 1988; Клишина, 1991; Benzecri, 1973; Clausen, 1998], канонический анализ [Интерпретация и анализ…, 1987; Thompson, 1984], конджойнт-анализ [Louvier, 1988], латентно-структурный анализ (ссылки см. в сноске [6] к части I), собственно алгоритмы оцифровки [Интерпретация и анализ…, 1987; Айвазян и др., 1983] и т.д. Сюда же с определенной оговоркой можно отнести методы многомерного шкалирования [Интерпретация и анализ…, 1987, гл. 8; Клигер и др., 1978, гл.4; Kruscal, Wish, 1978].

Эти методы, как известно, работают не с матрицами типа "объект-признак", а с матрицами близостей между шкалируемыми объектами; но интересующее нас положение остается в силе: предполагается, что респондент, так или иначе дающий оценку объектам, мыслит последние как точки в некотором пространстве восприятия, оси которого – непрерывные числовые переменные; задача же состоит в нахождении этих переменных (т.е. в

определении того, какова их суть, каковы их значения для каждого респондента). Сюда же можно отнести и многие известные методы

186

построения социологических индексов, например, известные способы одномерного шкалирования, связываемые обычно с именами Терстоуна, Лайкерта, Гуттмана. Перечисленные методы нами рассматриваться не будут.

Однако в рамках второго подхода находятся и некоторые методы другого рода, в том числе методы, позволяющие искать взаимодействия (CHAID) и измерять связь как между номинальными признаками в целом (Хи-квадрат), так и между отдельными группами альтернатив, отвечающих таким признакам (анализ фрагментов таблицы сопряженности). Эти методы будут подробно рассмотрены ниже, а CHAID будет сравнен с теми методами поиска взаимодействий, которые не опираются на существование упомянутой латентной переменной.

## 2.3. Анализ связей типа "признак-признак"

Для измерения связи между двумя номинальными признаками в литературе предлагается более сотни коэффициентов. Это является следствием того, что интересующее нас явление - указанную связь (еще раз подчеркнем, что мы говорим о статистической связи, хотя в действительности нас, как правило, интересуют соответствующие причинно-следственные отношения) – оказывается возможным формализовать по-разному. И каждому способу формализации отвечает свое понимание сути искомой связи, своя априорная модель того, что мы хотим изучить.

Мы не будем описывать все известные из литературы коэффициенты рассматриваемого характера. Коснемся лишь трех подходов к измерению парной связи между номинальными признаками. Эти подходы являются наиболее употребительными на практике. Надеемся, что их анализ, осуществленный ниже, заставит читателя "почувствовать" ту сложность социальной реальности, которая обусловливает возможность выделения в ней разных сторон, каждая из которых по-своему "представляет" изучаемое явление, по-своему формализуется.

187

### 2.3.1. Коэффициенты связи, основанные на критерии "хи-квадрат"

#### 2.3.1.1. Понимание отсутствия связи между признаками как их статистической независимости.

Приведем простой пример, иллюстрирующий рассматриваемый подход к пониманию связи между двумя номинальными признаками. Предположим, что перед нами стоит задача оценки того, зависит ли профессия респондента от его пола. Пусть наша анкета содержит соответствующие вопросы и в ней перечисляются пять вариантов профессий, закодированных цифрами от 1 до 5; для обозначения же мужчин и женщин используются коды 1 и 2 соответственно. Для краткости обозначим первый признак (т.е. признак, отвечающий вопросу о профессии респондента) через Y, а второй (отвечающий полу) - через X. Итак, наша задача состоит в том, чтобы определить, зависит ли Y от X.

Предположим, что исходная таблица сопряженности, вычисленная для каких-то 100 респондентов имеет вид:

*Таблица 8.*

**Пример таблицы сопряженности для двух независимых признаков**

| Профессия | Пол | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1 | 18 | 2 | 20 |
| 2 | 18 | 2 | 20 |
| 3 | 45 | 5 | 50 |
| 4 | 0 | 0 | 0 |
| 5 | 9 | 1 | 10 |
| Итого | 90 | 10 | 100 |

Вероятно, любой человек согласится, что в таком случае признаки можно считать независимыми, поскольку и мужчины, и женщины в равной степени выбирают ту или иную профессию: первая и вторая профессии пользуются одинаковой популярностью и у тех и у других; третью – выбирает половина мужчин, но и половина женщин; четвертую не любят ни те, ни другие и т.д. Итак, мы делаем вывод: независимость признаков означает пропорциональность столбцов (строк; с помощью несложиных арифметических выкладок можно показать, что пропорциональность столбцов эквивалентна пропорциональности строк) исходной частотной таблицы. Заметим, что в случае пропорциональности "внутренних" столбцов таблицы сопряженности, эти столбцы будут пропорциональны также и столбцу

маргинальных сумм по строкам. То же – и для случая пропорциональности строк они будут пропорциональны и строке маригинальных сумм по столбцам.

Приведенная частотная таблица получена эмпирическим путем, является результатом изучения выборочной совокупности респондентов. Вспомним, что в действительности нас интересует не выборка, а генеральная совокупность. Из математической статистики мы знаем, что выборочные данные никогда стопроцентно не отвечают “генеральным”. Любая, самая хорошая выборка всегда будет отражать генеральную совокупность лишь с некоторым приближением, любая закономерность будет содержать т.н. выборочную ошибку, случайную погрешность. Учитывая это, мы, вероятно, будем полагать, что, если столбцы выборочной таблицы сопряженности мало отличаются от пропорциональных, то такое отличие скорее всего объясняется именно выборочной погрешностью и вряд ли говорит о том, что в генеральной совокупности наши признаки связаны. Так мы проинтерпретируем, например, таблицу 9 (по сравнению с таблицей 8 в ней четыре частоты изменены на единицу) и, наверное, таблицу 10 (те же частоты изменены на две единицы). А как быть с таблицей 11?

Таблица 9.

**Первый пример таблицы сопряженности, частоты которой мало отличаются от ситуации независимости признаков**

| Профессия | Пол | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1 | 17 | 3 | 20 |
| 2 | 19 | 1 | 20 |
| 3 | 45 | 5 | 50 |
| 4 | 0 | 0 | 0 |
| 5 | 9 | 1 | 10 |
| Итого | 90 | 10 | 100 |

*Таблица 10.*

**Второй пример таблицы сопряженности, частоты которой сравнительно мало отличаются от ситуации независимости признаков**

| Профессия | Пол | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1 | 16 | 4 | 20 |

| | | | |
|---|---|---|---|
| 2 | 20 | 0 | 20 |
| 3 | 45 | 5 | 50 |
| 4 | 0 | 0 | 0 |
| 5 | 9 | 1 | 10 |
| Итого | 90 | 10 | 100 |

*Таблица 11.*

**Пример таблицы сопряженности, частоты которой значительно отличаются от ситуации независимости признаков**

| Профессия | Пол | | Итого |
|---|---|---|---|
| | 1 | 2 | |
| 1 | 15 | 5 | 20 |
| 2 | 20 | 0 | 20 |
| 3 | 46 | 4 | 50 |
| 4 | 0 | 0 | 0 |
| 5 | 9 | 1 | 10 |
| Итого | 90 | 10 | 100 |

Общая идея здесь ясна: сильное отклонение от пропорциональности заставляет нас сомневаться в отсутствии связи в генеральной совокупности, слабое отклонение говорит о том, что наша выборка не дает нам оснований для таких сомнений. Но насколько сильным должно быть указанное отклонение для того, чтобы описанные сомнения возникли?

Наука не дает точного ответа. Она предлагает нам лишь такой его вариант, который формулируется в вероятностных терминах. Этот ответ можно найти в математической статистике. Чтобы его воспринять, необходимо взглянуть на изучаемую связь, опираясь на своеобразное математико-статистическое видение мира. Опишем соответствующие рассуждения в следующем параграфе. Сразу скажем, что эти рассуждения типичны для математической статистики – речь идет об одной из основных решаемых ей задач – проверке статистической гипотезы.

#### 2.3.1.2. Функция "Хи-квадрат" и проверка на ее основе гипотезы об отсутствии связи

Предположим, что мы имеем две номинальных переменных, отвечающую им частотную таблицу типа 7 и хотим на основе ее анализа определить, имеется ли связь между переменными. Будем искать ответ на этот вопрос с помощью проверки статистической гипотезы о независимости признаков. Используя терминологию математической статистики, можно сказать, что речь пойдет о проверке нуль гипотезы $H_0$: “связь между рассматриваемыми переменными отсутствует”.

Далеко не для каждой интересующей социолога гипотезы математическая статистика предоставляет возможность ее проверки, не для каждой гипотезы разработана соответствующая теория. Но если упомянутая возможность существует, что соответствующая логика рассуждений сводится к следующему.

Допустим, что для какой-то статистической гипотезы $H_0$ разработана упомянутая теория и мы хотим эту гипотезу проверить. Математическая статистика предлагает некий критерий. Он представляет собой определенную числовую функцию **f** от наблюдаемых величин, например, рассчитанную на основе частот выборочной таблицы сопряженности: $\mathbf{f = f\ (n_{ij})}$ . Представим теперь, что в нашем распоряжении имеется много выборок, для каждой из которых мы можем вычислить значение этой функции. Распределение таких значений в предположении, что проверяемая гипотеза справедлива (для генеральной совокупности), хорошо изучено, т.е. известно, какова вероятность попадания каждого значения в любой интервал. Грубо говоря, это означает, что, если $H_0$ справедлива, то для каждого полученного для конкретной выборки значения **f** можно сказать, какова та вероятность, с которой мы могли на него “наткнуться”. Вычисляем значение $\mathbf{f}_{выб}$ критерия **f** для нашей единственной выборки. Находим вероятность $\mathbf{P(f}_{выб}\mathbf{)}$ этого значения.

Далее вступает в силу своеобразный принцип невозможности маловероятных событий: мы полагаем, что если вероятность какого-либо события очень мала, то это событие практически не может произойти. И если мы все же такое маловероятное событие встретили, то делаем из этого вывод, что вероятность определялась нами неправильно, что в действительности встреченное событие не маловероятно.

Наше событие состоит в том, что критерий принял то или иное значение. Если вероятность этого события (т.е. $\mathbf{P(f}_{выб}\mathbf{)}$) очень мала, то, в соответствии с приведенными рассуждениями, мы полагаем, что неправильно ее определили. Встает вопрос о том, что привело нас к ошибке. Вспоминаем, что мы находили вероятность в предположении справедливости проверяемой гипотезы. Именно это предположение и заставило нас считать вероятность

встреченного значения очень малой. Поскольку опыт дает основания полагать, что в действительности вероятность не столь мала, остается отвергнуть нашу $H_0$.

Если же вероятность $\mathbf{P(f_{выб})}$ достаточно велика для того, чтобы значение $\mathbf{f_{выб}}$ могло встретиться практически, то мы полагаем, что у нас нет оснований сомневаться в справедливости проверяемой гипотезы. Мы принимаем последнюю, считаем, что она справедлива для генеральной совокупности.

Таким образом, право именоваться критерием функция f обретает в силу того, что именно величина ее значения играет определяющую роль в выборе одной из двух альтернатив: принятия гипотезы $H_0$ или отвержения ее.

Остался нерешенным вопрос о том, где граница между “малой” и “достаточно большой” вероятностью? Эта граница должна быть равна такому значению вероятности, относительно которого мы могли бы считать, что событие с такой (или с меньшей) вероятностью практически не может случиться – “не может быть, потому, что не может быть никогда”. Это значение называют уровнем значимости принятия (отвержения) нуль-гипотезы и обозначают буквой $\alpha$. Обычно полагают, что $\alpha = 0,05$, либо $\alpha = 0,01$. Математическая статистика не дает нам правил определения $\alpha$. Установить уровень значимости может помочь только практика.Конечно, этот уровень должен обусловливаться реальной задачей, тем, насколько социально значимым может явиться принятие ложной или отвержение истинной гипотезы (процесс проверки статистических гипотез всегда сопряжен с тем, что мы рискуем совершить одну из упомянутых ошибок). Если большие затраты (материальные, либо духовные) связаны с отвержением гипотезы, то мы будем стремиться сделать $\alpha$ как можно меньше, чтобы была как можно меньше вероятность отвержения правильной нуль-гипотезы. Если же затраты сопряжены с принятием гипотезы, то имеет смысл $\alpha$ увеличить, чтобы уменьшить вероятность принятия ложной гипотезы.

Теперь рассмотрим конкретную интересующую нас нулевую гипотезу: гипотезу об отсутствии связи между двумя изучаемыми номинальными переменными. Функция, выступающая в качестве описанного выше статистического критерия носит название “хи-квадрат”, обозначается иногда как $\chi^2$ (Х - большое греческое “хи”; подчеркнем, что далее будет фигурировать малая буква с тем же названием; и надо различать понятия, стоящие за этими обозначениями, что не всегда делается в ориентированной на социолога литературе). Определяется этот критерий следующим образом:

$$\chi^2 = \sum_{i,j} \left[ \frac{(n_{ij}^{теор} - n_{ij}^{эмп})^2}{n_{ij}^{теор}} \right]$$

где $n_{ij}^{эмп}$ – наблюдаемая нами частота, стоящая на пересечении i -й строки и j -го столбца таблицы сопряженности (т.н. эмпирическая частота), а $n_{ij}^{теор}$ – та частота, которая стояла бы в той же клетке, если бы наши переменные были статистически независимы (т.е. та, которая отвечает пропорциональности столбцов (строк) таблицы сопряженности; она обычно называется теоретической, поскольку может быть найдена из теоретических соображений; иногда ее называют также ожидаемой частотой, поскольку действительно ее появление и ожидается при независимости переменных). Теоретическая частота обычно находится по формуле:

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

Приведем доказательство этой формулы. Сделаем это не для приобщения читателя к математике, а для демонстрации того, как необходимо воспринимать частоты при грамотном анализе таблицы сопряженности. Доказательство, о котором мы говорим, является очень простым, и использующиеся в процессе его проведения принципы входят в число тех знаний, которыми должен владеть каждый социолог, анализирующий эмпирические данные.

Итак, мы утверждаем, что теоретическая частота отвечает той ситуации, когда являются независимыми два события - то, что первый признак принимает значение i, и то, что второй признак принимает значение j. Независимость же двух событий означает, что вероятность их совместного осуществления равна произведению вероятностей осуществления каждого в отдельности. Вычислим соответствующие вероятности для интересующего нас случая. Представляется очевидным, что эти вероятности хорошо оцениваются (имеются в виду выборочные оценки вероятностей с помощью относительных частот) следующим образом:

$$P(X = i.Y = j) = \frac{n_{ij}}{n};\ P(X = i) = \frac{n_{i\bullet}}{n};\ P(Y = j) = \frac{n_{\bullet j}}{n}$$

Независимость наших событий означает справедливость соотношения:

$$P(X = i.Y = j) = P(X = i) \times P(Y = j)$$

или, учитывая введенные выше соотношения:

$$\frac{n_{ij}}{n} = \left( \frac{n_{i\bullet}}{n} \right) \times \left( \frac{n_{\bullet j}}{n} \right)$$

что легко преобразется в доказываемое соотношение (1). Перейдем к описанию того, как “работает” наш критерий “хи-квадрат”.

Представим себе, что мы организуем бесконечное количество выборок и для каждой из них вычисляем величину $\chi^2$. Образуется последовательность таких величин:

$\chi^2_{выб1}$, $\chi^2_{выб2}$, $\chi^2_{выб3}$, …

Очевидно, имеет смысл говорить об их распределении, т.е. об указании вероятности встречаемости каждого значения. В математической статистике доказано следующее положение: если наши признаки в генеральной совокупности независимы, то вычисленные для выборок значения $\chi^2$ приблизительно имеют хорошо изученное распределение, "имя" которого - $\chi^2$ ("хи-квадрат", здесь используется малое греческое "хи"). Приблизительность можно игнорировать (т.е. считать, что величины $\chi^2$ распределены в точности по закону $\chi^2$), если клетки тех выборочных частотных таблиц, на базе которых рассчитываются величины $\chi^2$, достаточно наполнены – обычно считают, что в каждой клетке должно быть по крайней мере 5 наблюдений. Будем считать, что это условие соблюдено.

Чтобы описание логики проверки нашей нуль-гипотезы стала более ясной, отметим, что отметим, что при отсутствии связи в генеральной совокупности среди выборочных $\chi^2$, конечно, будут преобладать значения, близкие к нулю, поскольку отсутствие связи означает равенство эмпирических и теоретических частот и, следовательно, равенство $\chi^2$ нулю. Большие значения $\chi^2$ будут встречаться сравнительно редко - именно они будут маловероятны. Поэтому можно сказать, что большое значение $\chi^2$ приводит нас к утверждению о наличии связи, малое – об ее отсутствии.

Теперь вспомним, что изученность распределения какой-либо случайной величины означает, что у нас имеется способ определения вероятности попадания каждого ее значения в любой заданный интервал – с помощью использования специальных вероятностных таблиц. Такие таблицы имеются и для распределения $\chi^2$. Правда, надо помнить, что такое распределение не одно. Имеется целое семейство подобных распределений. Вид каждого зависит от размеров используемых частотных таблиц. Точнее, этот вид определяется т.н. числом степеней свободы df (degree freedom) распределения, определяемым следующим образом:

$df = (r - 1) \times (c - 1)$.

Итак, если в генеральной совокупности признаки независимы, то, вычислив число степеней свободы для интересующей нас матрицы, мы можем найти по соответствующей таблице вероятность попадания произвольного значения $\chi^2$ в любой заданный интервал. Теперь

вспомним, что такое значение у нас одно – вычисленное для нашей единственной выборки. Обозначим его через $\chi^2_{выб}$. Описанная выше логика проверки статистической гипотезы превращается в следующее рассуждение.

Вычислим число степеней свободы df и зададимся некоторым уровнем значимости $\alpha$. Найдем по таблице распределения $\chi^2$ такое значение $\chi^2_{табл}$, называемое критическим значением критерия (иногда используется обозначение $\chi^2_{крит}$), для которого выполняется неравенство:

$P(\xi \geq i\,\chi^2_{табл}) = \alpha$

($\xi$ – обозначение случайной величины, имеющей распределение $\chi^2$ с рассматриваемым числом степеней свободы).

Если $\chi^2_{выб} < \chi^2_{табл}$ (т.е. вероятность появления $\chi^2_{выб}$ достаточно велика), то полагаем, что наши выборочные наблюдения не дают оснований сомневаться в том, что в генеральной совокупности признаки действительно независимы – ведь, “ткнув” в одну выборку, мы встретили значение $\chi^2$, которое действительно вполне могло встретиться при независимости. В таком случае мы полагаем, что у нас нет оснований отвергать нашу нуль-гипотезу и мы ее принимаем – считаем, что признаки независимы. Если же $\chi^2_{выб} \geq \chi^2_{табл}$ (т.е. вероятность появления $\chi^2_{выб}$ очень мала, т.е. меньше $\alpha$), то мы вправе засомневаться в нашем предположении о независимости – ведь мы “наткнулись” на такое событие, которое вроде бы не должно было встретиться при этом предположении. В таком случае мы отвергаем нашу нуль-гипотезу – полагаем, что признаки зависимы.

Итак, рассматриваемый критерий не гарантирует наличие связи, не измеряет ее величину. Он либо говорит о том, что эмпирия не дает оснований сомневаться в отсутствии связи, либо, напротив, дает повод для сомнений.

### 2.3.1.3. Нормировка значений функции "Хи-квадрат”.

Сами значения рассматриваемого критерия непригодны для оценки связи между признаками, поскольку они зависят от объема выборки и других обстоятельств, носящих, вообще говоря случайный характер по отношению к силе измеряемой связи (о некоторых обстоятельствах подобного рода пойдет речь ниже). Так, величина критерия, например, равная 30, может говорить о большой вероятности наличия связи, если в клетках исходной частотной

таблицы стоят величины порядка 10,20,30, и о ничтожной вероятности того же, если рассматриваемые частоты равны 1000, 2000, 3000 и т.д. В таких случаях возникает необходимость определенной нормировки найденного значения критерия – такого его преобразования, которое устранит описанную зависимость от случайных (для оценки связи) факторов.

Подчеркнем, что здесь речь идет о принципиальном моменте, часто возникающем при использовании в социологии разного рода статистических критериев, индексов и т.д. Всегда необходимо выяснять, не отражает ли используемый показатель что-либо случайное по отношению к изучаемому явлению и в случае наличия такого отражения осуществлять соответствующую нормировку показателя.

Принято нормировку, подобную описанной, осуществлять таким образом, чтобы нормированные коэффициенты изменялись либо от -1 до +1 (если имеет смысл противопоставление положительной и отрицательной направленности изучаемого с помощью рассматриваемого индекса явления, в нашем случае - связи), либо от 0 до 1 (если выделение положительной и отрицательной направленности явления содержательно бессмысленно).

Почеркнем, что приведение всех коэффициенту к одному и тому же интервалу является необходимым, но не достаточным условием, обеспечивающим возможность их сравнения. Если такого приведения не будет сделано, сравнение заведомо невозможно. Но и при его осуществлении сравнение тоже может оказаться бессмысленным. Об этом пойдет речь в п. 2.3.5.

Имеются разные подходы к требующейся нормировке. Наиболее известными являются такие, которые превращают критерий “Хи-квадрат” в известные коэффициенты, называемые обычно по именам впервые предложивших их авторов - Пирсона, Чупрова, Крамера. За этими коэффициентами утвердились постоянные обозначения, отвечающие первым буквам названных фамилий (коэффициент Чупрова отвечает немецкому tsch, коэффициент Крамера имеет два обозначения из-за известного различия букв, обозначающих звук “к” в разных языках):

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(c-1)(r-1)}}}$$

$$K(или C) = \sqrt{\frac{\chi^2}{n \times \min(c-1, r-1)}}$$

Опишем некоторые свойства этих коэффициентов. Начнем с тех, которые обычно оговариваются в литературе.

Все коэффициенты изменяются от 0 до 1 и равны нулю в случае полной независимости признаков (в описанном выше смысле). Как и критерий “хи-квадрат”, эти показатели являются симметричными относительно наших признаков: с их помощью нельзя выделить зависимую и независимую переменную, на основе их анализа нельзя говорить о том, какая переменная на какую "влияет".

Обычно в качестве недостатка коэффициента Пирсона Р (предложенного в литературе первым) упоминается зависимость его максимальной величины от размера таблицы (максимум Р достигается при с=r, но величина максимального значения изменяется с изменением числа категорий: при с=3 значение Р не может быть больше 0,8, при с=5 максимальное значение Р равно 0,89 и т.д. [ Интерпретация и анализ ..., 1987. С.31]). Естественно, это приводит к возникновению трудностей при сравнении таблиц разного размера.

Отметим следующий немаловажный факт, очень редко рассматривающийся в ориентированной на социолога литературе.

Многие свойства рассматриваемых коэффициентов доказываются лишь при условии выполнения одного не всегда приемлемого для социологии предположения, состоящего в том, что за каждым нашим номинальным признаком "стоит" некая латентная (скрытая) непрерывная количественная (числовая) переменная.

Сделаем небольшое отступление по поводу используемых терминов. Все три определения к термину "переменная" требуют пояснения. Термин "латентная" употребляется здесь несколько условно. Обычно (в теории социологического измерения, например, в факторном, латентно-структурном анализе, многомерном шкалировании) под латентной переменной понимают признак, значения которого вообще не поддаются непосредственному измерению (например, путем прямого обращения к респонденту). Значения же нашей переменной мы измеряем самым непосредственным образом. Но получаем при этом номинальную шкалу, хотя и предполагаем, что между отвечающими этим значениям свойствами реальных объектов существуют отношения, достаточно сложные для того, чтобы можно было говорить об использовании интервальной шкалы (о соотношении между "богатством" реальных отношений между эмпирическими объектами и типом шкал, использующихся при шкалировании этих объектов, см., например [Клигер и др., 1978; Толстова, 1998]).

Термин "непрерывная" здесь употребляется в том смысле, что в качестве значения этой переменной может выступать любое рациональное число.

"Количественной" мы, в соответствии с традицией, называем переменную, значения которой получены по шкале, тип которой не ниже типа интервальной шкалы (о нашем отношении к подобному использованию терминов "качественный - количественный" уже шла речь в п.4.3 части I). Можно показать, что для таких шкал любое рациональное число может в принципе оказаться шкальным значением какого-либо объекта. Поэтому термины "количественный" и "непрерывный" часто употребляются как синонимы.)

Итак, мы полагаем, что каждый номинальный признак получен из некоторого количественного в результате произвольного разбиения диапазона его изменения на интервалы, количество которых равно числу значений нашей номинальной переменной. И, задавая респонденту интересующий нас вопрос в анкете, мы как бы принуждаем его разбить весь диапазон изменения рассматриваемой переменной на интервалы и указать, в каком из этих интервалов, по его мнению, находится оцениваемый объект. Внутри каждого интервала значения переменной становятся неразличимыми, между интервалами же определены лишь отношения совпадения – несовпадения (основное свойство номинальной шкалы). Когда исследователь имеет дело с двумя переменными такого рода (например, когда нас интересуют парные связи) то обычно предполагается еще и нормальность соответствующего двумерного распределения.

Именно таких предположений придерживался Пирсон, когда в начале века вводил свой коэффициент. Он доказал, что Р равно тому предельному значению обычного коэффициента корреляции между латентными переменными, к которому этот коэффициент стремится при безграничном увеличении количества градаций рассматриваемых признаков. Ясно, что без указанного предположения было бы совершенно неясно, как подобное свойство коэффициента Р можно проинтерпретировать.

Для исправления указанного недостатка коэффициента Пирсона (зависимости его максимально возможного значения от размеров таблицы сопряженности) Чупров ввел коэффициент Т, названный его именем. Но и Т достигает единицы лишь при $c=r$, и не достигает при $c \neq r$. Может достигать единицы независимо от вида таблицы коэффициент Крамера К. Для квадратных таблиц коэффициенты Крамера и Чупрова совпадают, в остальных случаях К >Т.

Мы перечислили те свойства рассматриваемых коэффициентов, которые часто упоминаются в литературе. Из редко упоминающихся свойств можно упомянуть еще один

свойственный всем коэффициентам недостаток – зависимость их величины от соотношений маргинальных частот анализируемой таблицы сопряженности (подчеркнем очень важный момент – вычисляя теоретические частоты, мы пользуемся маргинальными суммами, полагая, что имеем дело с их “генеральными” значениями, что, вообще говоря, не всегда отвечает реальности).

О том, как можно измерять связь между номинальными признаками с помощью критерия “Хи-квадрат”, можно прочесть в работах [Елисеева, 1982; Елисеева, Рукавишников, 1977, с.82-89; Интерпретация и анализ ..., 1987, с.31-32; Лакутин, Толстова, 1990; Паниотто, Максименко, 1982, с.65-84; Рабочая книга социолога, 1983, с.169-172, 190 (с учетом того, что на с, 169 речь идет о таких теоретических частотах, которые являются частотами таблицы сопряженности, отвечающей случаю статистической независимости рассматриваемых номинальных переменных); Статистические методы ..., 1979, с.117-120; Толстова, 1990а, с.54-57]

Перейдем к описанию таких коэффициентов парной связи, которые основаны на других априорных моделях, на другом понимании сути этой связи.

### 2.3.2. Коэффициенты связи, основанные на моделях прогноза

#### 2.3.2.1. Выражение представлений о связи через прогноз

Включение понятия прогноза в представление о связи между номинальными признаками представляется разумным: наверное, трудно возражать против того, чтобы признаки считались связанными, если знание значения одного признака позволяет улучшить прогноз значения другого. Поясним это на гипотетическом примере, который ниже мы будем неоднократно “эксплуатировать”. Заодно уточним только что сформулированное суждение.

201

Предположим, что мы изучаем жителей некоторого крупного города N от 20 лет и старше и что нас интересует связь между признаком “возраст”, рассматриваемым нами как номинальный и дихотомическим признаком со значениями “студент” – “не студент”.

(Напомним два принципиальных для социологии момента. Во-первых, определение типа шкалы для таких, казалось бы, “понятных” признаков, как возраст, далеко не всегда является ясным делом; причиной тому служит то, что их значения, как правило, интересуют исследователя не сами по себе, а лишь как показатели некоторых латентных переменных. Во-вторых, здесь мы отвлекаемся от сложной проблемы разбиения диапазона изменения

непрерывного признака – предполагаем, что это сделано каким-либо адекватным решаемой задаче образом.)

Предположим, что распределение изучаемой совокупности по возрасту приблизительно равномерно, например, такое, какое изображено на рис. 14.

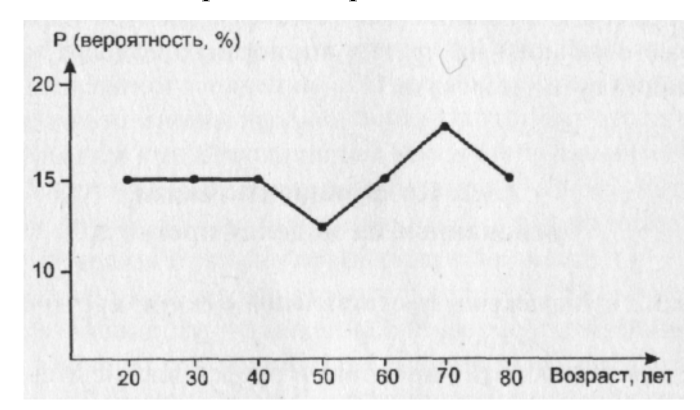



**Рис.14. Гипотетическое распределение по возрасту жителей города N старше 20 лет**

Интуитивно ясно, что в такой ситуации мы вряд ли сможем хорошо прогнозировать возраст респондента. Выбрав наугад (случайным образом) произвольного человека, мы примерно с

202

одинаковой степенью уверенности можем полагать, что он имеет любой возраст: вероятность "наткнуться" на 20-летнего юношу такая же, как и на 80-летнего старика (подчеркнем своеобразие понимания нами термина "прогноз" - речь идет просто о том, что мы можем сказать о значении возраста для случайно выбранного респондента).

Другое дело, если мы рассмотрим только студентов. Ясно, что их распределение по возрасту будет резко отличаться от общего. Например, будет иметь вид, изображенный на рис. 15.

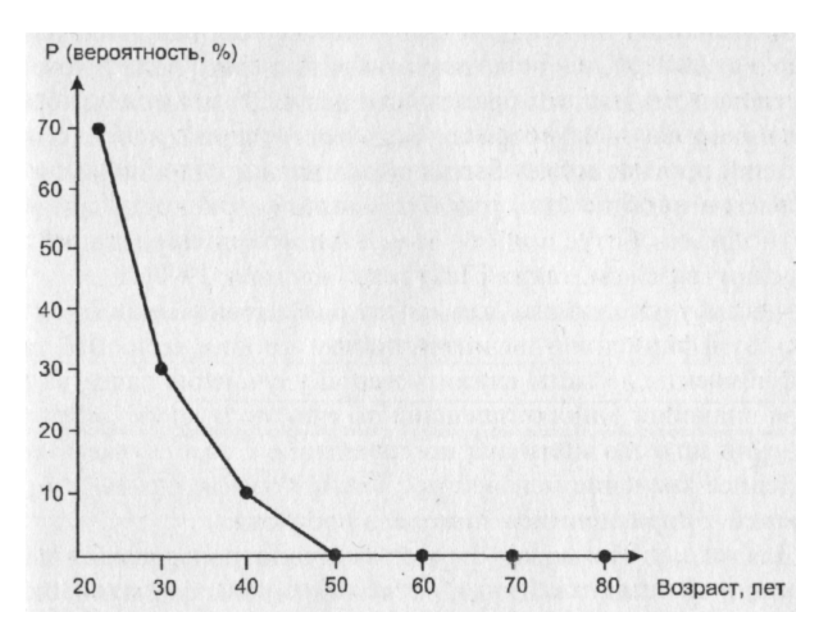



**Рис. 15. Гипотетическое распределение по возрасту студентов города N старше 20 лет**

Ясно, что теперь, случайным образом отобрав человека (студента), мы с уверенностью 90% (90 = 70 + 20) будем полагать, что его возраст не превысит 30 лет, вероятность же "попасть" на человека старше 40 лет практически равна нулю.

Итак, фиксировав значение "студент" второго рассматриваемого нами признака, мы явно улучшили возможность прогноза

203

возраста жителей города. Наверное, на основе этого было бы разумно сделать вывод о наличии связи между признаком "возраст" и признаком "быть студентом". Подчеркнем, что для того, чтобы сделать этот вывод, мы *сравнили* безусловное распределение признака "возраст" (рис. 14) с его условным распределением (рис. 15), когда условие состоит в фиксации значения "студент" второго признака. Возможность хорошего прогноза на основе знания условного распределения сама по себе (без ее сравнения с возможностью прогноза по безусловному распределению) ни о какой связи еще не говорит. Так, изучая только студентов, мы не можем говорить о связи пола и возраста на основе того, .что, отобрав только девушек, мы можем хорошо прогнозировать их возраст. Ведь, всего вероятнее, столь же хороший прогноз может быть осуществлен и для юношей, и для студентов вообще (т.е. для безусловного распределения). О соотношении безусловного и условного распределений при изучении связей см. также [Лакутин, Толстова, 1990].

Итак, будем считать, что смысл рассматриваемых (прогнозных) коэффициентов на интуитивном уровне ясен. Все такие коэффициенты должны служить мерой улучшения качества прогноза значения одного признака за счет получении сведений о значении другого признака по сравнению с тем случаем, когда последнее значение неизвестно. Такие коэффициенты и будем называть опирающимися на модель прогноза.

Для того, чтобы можно было практически пользоваться высказанными предположениями, необходимо их формализовать. Другими словами, необходимо четко понять, что такое прогноз и как именно на основе частотной таблицы мы можем судить о различии возможности прогноза для соответствующих условных и безусловных распределений. Формализация может быть разной. И, в первую очередь, неоднозначно может пониматься сам термин "прогноз". Те известные коэффициенты связи, которые мы намереваемся рассмотреть, отличаются друг от друга как раз способом формализации этого понятия. Но прежде, чем переходить к описанию некоторых прогнозных коэффициентов, напомним, что проблема формализации содержательных

204

представлений о "прогнозной" связи, вообще говоря, не исчерпывается рассуждениями о понимании прогноза и оценке его качества. Отметим также следующие три немаловажные момента.

Во-первых, глобальные коэффициенты связи по существу являются "усреднениями" всевозможных локальных коэффициентов. И подобные "усреднения" могут пониматься по-разному, выражаться разными формулами. Это также обусловливает наличие разных коэффициентов связи.

Во-вторых, возможность осуществления прогноза значений одного признака по значениям другого существенно зависит от того, значения какого признака прогнозируются. Скажем, значения первого могут хорошо прогнозироваться по значениям второго, а значения второго по значениям первого - очень плохо. Приведем простой, несколько утрированный пример. Пусть частотное распределение значений двух признаков имеет вид, представленный в табл. 12.

*Таблица 12*

**Таблица сопряженности, иллюстрирующая несимметричность понятия "прогноз"**

| X | Y |
|---|---|

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0 | 0 | 10 |
| 2 | 0 | 0 | 10 |
| 3 | 0 | 20 | 0 |

Ясно, что по значению X мы легко предсказываем значение Y. Обратное же не имеет места: если признак Y равен 3, то X с одинаковым успехом (с равной вероятностью) может принимать значения 1 или 2. В таком случае возникает вопрос о построении коэффициентов, не симметричных относительно рассматриваемых признаков или, как говорят, коэффициентов, отражающих направленную связь – скажем, говорящих о том, появляется ли у нас новая информация о втором признаке при фиксации значения первого, но ничего не говорящих об обратной зависимости.

Актуальной является задача усреднения таких направленных коэффициентов для оценки ненаправленной связи. Обоснование

205

соответствующей необходимости - примерно такое же, как обоснование необходимости использования глобальных коэффициентов наряду с локальными: с одной стороны. не имея коэффициентов направленной связи, мы можем упустить, не заметить важные причинно-следственные отношения, но, с другой – когда направленные связи не очень значимы, мы можем “за деревьями” не увидеть леса” – не уловить того, что, хотя каждая направленная связь не очень велика, в целом нельзя игнорировать взаимодействие рассматриваемых признаков.

О терминах: когда говорят о прогнозе значения признака Y по признаку X, то X называют независимой переменной, а Y – зависимой.

Перейдем к описанию наиболее известных коэффициентов, основанных на моделях прогноза.

### 2.3.2.2. Коэффициенты, основанные на модальном прогнозе

Формализуем понятие прогноза следующим образом. Выбирая произвольный объект и зная распределение рассматриваемого признака (условное или безусловное), считаем, что для выбранного объекта этот признак принимает то значение, которое имеет максимальную вероятность, встречается с максимальной частотой (т.е. модальное значение). Такой прогноз называется модальным. Чтобы стал ясен содержательный смысл рассматриваемого прогноза,

приведем формулы соответствующих коэффициентов. Но сначала отметим, что таких коэффициентов три: два отражают возможные направленные связи, а третий является их усреднением. Эти коэффициенты обычно обозначаются буквами λ с индексами: $\lambda_r$ – отражающий “влияние” строкового признака на столбцовый; $\lambda_c$ – отражающий "влияние" столбцового признака на строковый, λ – усредненный коэффициент.

Рассмотрим формулу для $\lambda_r$, (для $\lambda_c$ рассуждения совершенно аналогичны). Будем использовать те же обозначения, которые были задействованы выше.

206

$$\lambda_r = \frac{\sum_{i=1}^{r} \max_i n_{ij} - \max_j n_{\bullet j}}{n - \max_j n_{\bullet j}} \quad (2)$$

Выражение $\max_i n_{ij}$ означает наибольшую частоту в i - й строке.

Выражение $\max_j n_{\cdot j}$ – наибольшую столбцовую маргинальную частоту.

Поясним смысл формулы (2) на примере. Пусть частотная таблица имеет вид:

*Таблица 13.*

**Пример частотной таблицы, использованный для расчета коэффициента $\lambda_r$**

| X | Y | | | Итого |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1 | 0 | 20 | 30 | 50 |
| 2 | 5 | 15 | 30 | 50 |
| 3 | 40 | 5 | 5 | 50 |
| Итого | 45 | 40 | 65 | 150 |

Наибольшая частота в первой строке матрицы равна 30, во второй – тоже 30, в третьей – 40. Максимальный маргинал по столбцам – 65. Общее количество объектов в выборе – 150. Значит, имеет место равенство:

$$\lambda_r = \frac{(30+30+40)-65}{150-65} = 0,41$$

Рассмотрим безусловное распределение признака Y. Отвечающие ему частоты – это маргиналы по столбцам рассматриваемой матрицы: 45, 40, 65. Модальная частота – 65. Значит, выбрав случайным образом какой-либо объект, мы, прогнозируя

207

для него значение Y, в соответствии с нашими представлениями о прогнозе, должны сказать, что упомянутое значение равно 3 (именно это значение является модой). Ясно, что, поступая так и перебирая последовательно всех респондентов, мы дадим правильный прогноз в 65 случаях и ошибемся в (150 - 65) случаях (заметим, что доля (вероятность) ошибки будет равна $\frac{150-65}{150}$. Именно эта разность стоит в знаменателе нашей формулы.

Итак, для безусловного распределения качество нашего прогноза можно оценить с помощью величины (150 - 65). Улучшится ли прогноз при переходе к условным распределениям того же признака? Попытаемся ответить на этот вопрос.

Пусть X = 1. Соответствующее условное распределение Y определяется частотами первой строки нашей матрицы: числами 0, 20, 30. Значит, перебирая 50 респондентов с первым значением X, и делая для каждого прогноз в соответствии с нашими правилами, мы не ошибемся в 30 случаях. При X = 2 количество верных предположений тоже будет равно 30. При X=3 – 40. Общее количество правильных прогнозов во всех условных распределениях будет равно (30+30+40). По сравнению с “безусловным” случаем оно возрастет на ((30+30+40) - 65) единиц. А это – числитель выражения для $\lambda_r$.

Итак, в числителе формулы (2) отражена величина того суммарного прироста количества правильных прогнозов, который возникает за счет перехода от перебора объектов, “сваленных в одну кучу” (“куча” отвечает безусловному распределению), к перебору последовательно по “слоям” (отвечающим условным распределениям). Эта величина отражает суть коэффициента. Знаменатель же формулы (2) использован для нормировки (знаменатель равен значению числителя, получающемуся, когда суммарный прогноз прогноз по условным распределениям будет стопроцентным). Потребность в таковой возникает в силу тех же причин, которые были обсуждены нами при рассмотрении критерия “хи-квадрат”: без нормировки величина коэффициента будет зависеть от размера выборки, значений конкретных частот и т.д.

208

Теперь, чтобы закончить вопрос о том, как в рассматриваемом случае формализуются естественные представления о связи, необходимо затронуть проблему “усреднения” всевозможных связей типа “альтернатива-альтернатива”. Способ усреднения очевиден. Он как бы двуступенчат. Рассматривая какое-либо из наших условных распределений, мы говорим о прогнозе, учитывая сразу все возможные значения Y, не анализируя отдельно, насколько

зафиксированное значение X может быть связано с тем или иным значением Y (в п. 2.3.2.3 мы увидим, как такая связь может быть прослежена).

Переходя к общей формуле, мы суммируем показатели качества прогноза для всех условных распределений, игнорируя то, что для одного значения X этот прогноз может быть хорошим, а для другого – плохим.

В заключение обсуждения вопроса о $\lambda_r$ опишем некоторые его свойства.

Имеют место неравенства: $0 \leq \lambda_r \leq 1$. Коэффициент приближается к 1 по мере того, как в каждой строке объекты все более концентрируются в одной клетке, т.е. прогноз значения Y для условных распределений становится все лучше. Нетрудно проверить, что $\lambda_r = 1$, если

$$\sum_{i=1}^{r} \max_j n_{ij} = n \, ,$$

и что это, в свою очередь, может быть верным лишь в случае, когда в каждой строке частотной таблицы существует только одна отличная от нуля частота, т.е. когда по значению признака X мы можем однозначно судить о значении признака Y (но не обратно!).Чем ближе значение $\lambda_r$ к 1, тем лучше такое предсказание и сильнее связь (в рассматриваемом понимании) между переменными.

$\lambda_r = 0$, если максимальные частоты в строках приходятся на один и тот же столбец. Это имеет место даже в том случае, если все остальные элменты частотной таблицы близки к нулю, т.е. если фактически имеется “хорошая” связь (а отнюдь не отсутствие связи,

209

как это должно было бы быть для нулевого значения хорошего коэффициента связи). И это является существенным недостатком рассматриваемого коэффициента.

Как мы уже отмечали, все приведенные рассуждения справедливы и для коэффициента $\lambda$, служащего показателем связи, если зависимая и независимая переменные меняются местами, и вычисляющегося по формуле:

$$\lambda_c = \frac{\sum_{j=1}^{c} \max_j n_{ij} - \max_i n_{i\bullet}}{n - \max_i n_{i\bullet}}$$

Для измерения по тому же принципу ненаправленной связи показатели рассматриваемых направленных связей усредняются. Это делается разными способами. Самый простой:

$$\lambda = \frac{\lambda_r + \lambda_c}{2}$$

Итак, подведем итог обсуждению рассмотренных коэффициентов. Правила их построения определяют отвечающее модальному классу значение зависимого признака (Y) как оценку этого значения для произвольно взятого объекта. Если оценка делается без знания значения независимого признака (X), то значением, предсказываемым для всех объектов, является модальное значение безусловного распределения зависимого. Если же оценка делается на основе знания значения X, то прогноз осуществляется отдельно для объектов, обладающих этим значением, на основе выявления моды соответствующего условного распределения Y. Величина $\lambda_г$ ($\lambda_с$) говорит об уменьшении (за счет осуществения перехода от безусловного распределения к набору условных) ошибки осуществленного с единичной вероятностью предсказания о том, что объект обладает модальным значением Y.

Приведем несколько утрированный пример. Рассмотрим, как может измеряться связь между национальностью (X) и цветом

210

волос (Y). Предположим, что Вы являетесь продавцом косметики и Вам для того, чтобы заранее подготовиться к общению с покупателем, желательно заранее знать цвет его волос. Представим себе, что вы арендовали помещение в вузе и к вам в комнату по очереди (в случайном порядке) входят за покупкой студенты. Допустим также, что Вы знаете безусловное распределение всех студентов рассматриваемого вуза по цвету волос, и в соответствии с этим распределением количество блондинов, брюнетов и шатенов примерно одинаково, но шатенов несколько больше, чем остальных. Вы пользуетесь правилом: перед входом покупателя приготавливаете товар, рассчитанный на модальное значение признака “цвет волос” (в нашем случае – на шатенов).

Теперь представим себе две ситуации.

В первой Вы ничего не знаете о национальности входящего к вам студента. Наверное, в таком случае, приготовив товар для шатенов, Вы в почти двух третях возможных случаев совершите ошибку: к Вам с одинаково вероятностью в любой момент может войти и блондин, и брюнет, и шатен. Торговля заведомо будет неэффективной.

А во второй ситуации Вы сумели организовать дело так, что сначала к Вам по очереди (снова в случайном порядке) входят учащиеся в вузе китайцы, затем - финны, потом - русские. Очевидно, эффективность Вашей торговли возрастет: зная, что сегодня к Вам придут китайцы, Вы готовите товар, рассчитанный только на брюнетов, если придут финны - на блондинов, если русские - на шатенов. Конечно, Вы и тут будете ошибаться, но уже в гораздо меньшей степени,

чем раньше. Другими словами, Ваш прогноз улучшится. А это и означает наличие связи между национальностью и цветом волос. Чем в большей мере прогноз улучшился, тем сильнее связь.

Описанный прогноз называют модальным, или оптимальным. Коэффициенты чаще всего называют коэффициентами Гуттмана [Интерпретация и анализ ..., 1987; Статистические методы ..., 1979], Гудмена [Паниотто, Максименко, 1982] или λ-коэффициентами [Рабочая книга, 1983].

211

### 2.3.2.3. Общее представление о пропорциональном прогнозе

Представленное понимание прогноза не является единственно возможным. Более того, его нельзя признать наилучшим. Прогноз здесь очень груб, приблизителен. Используя достижения теории вероятностей, к определению понятия прогноза можно подойти более тонко. Опишем еще один подход. На нем тоже базируется целый ряд известных коэффициентов связи (например, коэффициент Валлиса [Интерпретация и анализ ..., 1987; Статистические методы ..., 1979]). Принцип их “действия” по существу является тем же, что и принцип λ-коэффициентов. Отличие состоит только в понимании процедуры прогноза. Мы не будем эти коэффициенты описывать, поскольку такое описание требует использования довольно сложных формул, но ничего не дает принципиально нового для понимания отражаемой с помощью этих коэффициентов связи.

Итак, что же такое пропорциональный прогноз? Опишем его суть с помощью примера.

Предположим, что мы имеем дело с частотной табл. 13. Рассмотрим безусловное распределение Y. Обратимся к схематичному изображению ситуации в терминах столь часто фигурирующих в литературе по теории вероятностей урн и заполняющих их шаров. Возьмем 150 шаров, на 45 из них напишем цифру 1, на 40 - цифру 2, на 65 - цифру 3 и погрузим все шары в урну, перемешав их. Правило прогноза выглядит очень просто: берем случайного респондента, опускаем руку в урну и вытаскиваем тот шар, который случайно же нам попался. То, что на нем написано, и будет прогнозным значением признака Y для выбранного респондента. Аналогичным образом поступаем и для каждого условного распределения. Конечно, реализовать такой подход можно и без шаров с урнами, но суть должна сохраниться: то, что чаще встречается в исходной совокупности, должно чаще попадаться в наши руки при вытаскивании шаров. К примеру, в соответствии с первым условным распределением (X=1,

первая строка частотной таблицы), у нас отсутствуют респонденты, для которых Y = 1. Не будут попадаться нам и шары с единицей,

212

поскольку количество таких шаров равно 0. В соответствии с третьим распределением (X=3) значения 2 и 3 признака Y встречаются одинаково часто и в 8 раз реже значения 1. И вероятность встречаемости шаров с цифрами 2 и 3 будет одинаковой и в 8 раз меньше вероятности встречаемости шара с 1.

Описанный прогноз называется пропорциональным. Хотя соответствующее правило на первый взгляд, довольно сложно, оно позволяет предсказывать значение зависимого признака с большей надежностью, чем правило модального прогноза. Это часто используется в самых разных прогнозных алгоритмах.

### 2.3.3. Коэффициенты связи, основанные на понятии энтропии

Семейство коэффициентов, к рассмотрению которых мы переходим, основаны на такой модели связи, которая очень близка по своему содержательному смыслу к прогнозным моделям. В основе этих коэффициентов также лежит сравнение безусловного распределения с условными (условие - фиксация значения независимосго признака X). Но сравнение это ведется не с точки зрения того, насколько при переходе от безусловного распределения к условным меняется качество возможного прогноза, а с точки зрения изучения изменения степени неопределенности рассматриваемых распределений. Здесь мы, как и в п. 1.3.5, вступаем в область теории информации и будем использовать ее терминологию.

#### 2.3.3.1. Условная и многомерная энтропия

Вернемся к рассмотренному нами в п. 1.3.5 раздела 1 понятию энтропии.

По аналогии с энтропией распределения одного признака, определяется энтропия двумерного распределения:

$$H(X,Y) = -\sum_{i,j} P(X = i.Y = j) \times \log P(X = i.Y = j)$$

213

Точка внутри скобок означает конъюнкцию соответствующих событий, одновременной их выполнение. Если ввести обозначения, аналогичные использованным выше: $P_{ij} = P(X = i.Y = j)$, то же соотношение запишется в виде:

$$H(X,Y) = -\sum_{i,j} P_{ij} \times \log P_{ij}$$

Точно так же можно определить энтропию любого многомерного распределения.

Необходимо дать определение еще одного очень важного для нас понятия – т.н. условной энтропии:

$$H(Y/X) = -\sum_{i} P_i \cdot H(Y/X = i) = \sum_{i} P_i \sum_{j} P(Y = j/X = i) \times \log P(Y = j/X = i) \qquad (3)$$

Можно доказать следующие свойства энтропии.

H (X,X) = H (X); H (X,Y) = H (X) + H (Y/X); H (X,Y) ≤ H (X) + H (Y);

равенство в последнем соотношении появляется только тогда, когда X и Y статистически независимы, т.е. когда выполняется уже обсужденное нами соотношение: $P_{ij} = P_i \times P_j$..

В определенном смысле противоположным понятию энтропии является понятие информации, к рассмотрению которого мы переходим.

(Отметим, что говоря об информации в сочетании с энтропией, мы вступаем в сферу мощного научного направления – теории информации. Решающим этапом в становлении этой теории явилась публикация ряда работ К.Шеннона)

Приобретение информации сопровождается уменьшением неопределенности, поэтому количество информации можно измерять количеством исчезнувшей неопределенности, т.е. степенью уменьшения энтропии. Ниже речь пойдет об информации, содержащейся в одном признаке (случайной величине) относительно другого признака. Поясним смысл этого понятия более

214

подробно, по существу используя другой язык для описания того же, о чем шла речь выше [Яглом, Яглом, 1980. С. 78].

Вернемся к величине H(Y), характеризующей степень неопределенности распределения Y или, говоря несколько иначе, степень неопределенности опыта, состоящего в том, что мы случайным образом отбираем некоторый объект и измеряем для него величину Y.

Если H(Y)=0, то исход опыта заранее известен. Большее или меньшее значение H(Y) означает большую или меньшую проблематичность результата опыта. Измерение признака X, предшествующее нашему опыту по измерению Y, может уменьшить количество возможных

исходов опыта и тем самым уменьшить степень его неопределенности. Для того, чтобы результат измерения X мог сказаться на опыте, состоящем в измерении Y, необходимо, чтобы упомянутый результат не был известен заранее. Значит, измерение X можно рассматривать как некий вспомогательный опыт, также имеющий несколько возможных исходов. Тот факт, что измерение X уменьшает степень неопределенности Y, находит свое отражение в том, что условная энтропия опыта, состоящего в измерении Y, при условии измерения X оказывается меньше (точнее, не больше) первоначальной энтропии того же опыта. При этом, если измерение Y не зависит от измерения X, то сведения об X не уменьшают энтропию Y, т.е. H(Y/X) = H (Y). Если же результат измерения X полностью определяет последующее измерение Y, то энтропия Y уменьшается до нуля:

H(Y/X) = 0.

Таким образом, разность

I(X,Y) = H(Y) – H(Y/X) (4)

указывает, насколько осуществление опыта по измерению X уменьшает неопределенность Y, т.е. сколько нового мы узнаем об Y, произведя измерение X. Эту разность называют *количеством информации* относительно Y, содержащейся в X (в научный обиход термин был введен Шенноном).

215

Приведенные рассуждения о смысле понятия информации очевидным образом отвечают описанной выше логике сравнения безусловного и условных распределений Y. В основе всех информационных мер связи (а о них пойдет речь ниже) лежит та разность, которая стоит в правой части равенства (4). Но именно эта разность и говорит о различии упомянутых распределений. Нетрудно понять и то, каким образом здесь происходит усреднение рассматриваемых характеристик всех условных распределений (напомним, что в качестве характеристики распределения у нас выступает его неопределенность, энтропия). По самому своему определению (см. соотношение (3)) выражение H(Y/X) есть взвешенная сумма всех условных энтропий (каждому значению признака X отвечает своя условная энтропия Y:

$$\sum_j P(Y = j / X = i) \times \log P(Y = j / X = i)$$

причем каждое слагаемое берется с весом, равным вероятности появления соответствующего условного распределения, т.е. вероятности $P_i$ . Другими словами, можно сделать вывод, что для выборки величина H(Y/X) - это обычное среднее взвешенное значение условных энтропий.

О возможных способах нормировки разности (H(Y) – H(Y/X)) пойдет речь далее, поскольку рассматриемые ниже коэффициенты именно этой нормировкой фактически и отличаются друг от друга.

В заключение настоящего параграфа опишем некоторые свойства информации.

I(X,Y) – функция, симметричная относительно аргументов, поскольку, как нетрудно показать, имеет место соотношение:

I(X,Y) = H(X) + H(Y) – H(X,Y),

а функция H(X, Y) симметрична по самому своему определению. Другими словами, количество информации, содержащейся в X относительно Y, равно количеству информации в Y относительно X, т.е. соотношение (4) эквивалентно соотношению

216

I(X,Y) = H(X) – H(X/Y),

Перейдем к описанию мер связи, основанных на понятии энтропии.

### 2.3.3.2. Смысл энтропийных коэффициентов связи. Их формальное выражение

Поскольку понятие энтропии является как бы обратной стороной понятия информации, то энтропийные коэффициенты в литературе нередко называют информационными. Мы эти два термина будем использовать как синонимы.

Переходя к обсуждению конкретных информационных мер связи, прежде всего отметим, что в качестве такой меры может служить I(X, Y). Как мы уже отметили, это - симметричная (значит, - ненаправленная) мера. Из приведенных выше свойств энтропии следуют следующие свойства названной меры:

I(X, Y) ≥ 0,

где равенство достигается тогда и только тогда, когда X и Y статистически независимы и

I(X, X) = H(X).

Широко известны и направленные меры связи:

$$C_{X/Y} = \frac{I(X,Y)}{H(X)} \text{ и } C_{Y/X} = \frac{I(Y,X)}{H(Y)}$$

Первый из этих коэффициентов можно интерпретировать как относительное приращение информации об X, возникающее за счет знания Y [Миркин, 1980. С. 103]. Относительность

возникает в результате соотнесения такого приращения с первоначальной неопределенностью распределения X. Аналогично интерпретируется и второй коэффициент.

217

Коэффициенты С называют асимметричными коэффициентами неопределенности, коэффициентами нормированной информации [Елисеева, Рукавишников,1977. С. 91]. Нетрудно проверить справедливость следующих соотношений [Елисеева, Рукавишников,1977; Статистические методы ..., 1979]:

$0 \le C_{X/Y} \le 1$;

$C_{X/Y} = 0$ если и только если переменные X и Y независимы; $C_{X/Y} = 1$, если и только если X однозначно определяется значением Y (т.е. если можно говорить о детерминистской зависимости X от Y; о том, что мера разнообразия X определяется мерой разнообразия Y единственным образом, т.е. о полной связи).

Ясно, что аналогичными свойствами обладает и коэффициент $C_{Y/X}$.

Соответствующий симметризованный коэффициент нормированной информации вводится следующим образом [Елисеева, Рукавишников,1977. С. 95]:

$$R(Y,X) = \frac{I(X,Y)}{0{,}5(H(X)+H(Y))}$$

Часто используется также коэффициент Райского:

$$R(Y,X) = \frac{I(X,Y)}{H(X,Y)}$$

Нетрудно проверить, что он обладает свойствами, аналогичными сформулированным выше свойствам коэффициентов С: заключен в интервале от 0 до 1, в 0 обращается тогда и только тогда, когда признаки статистически независимы, а в 1 – тогда и только тогда, когда признаки полностью детерминируют друг друга.

Введенные информационные меры связи во многом похожи на обычный коэффициент корреляции. Но они имеют одно преимущество перед последним: из того, что коэффициент

218

корреляции равен 0, вообще говоря, не следует статистическая независимость рассматриваемы признаков, а из равенства 0 рассмотренных информационных мер связи – следует.

Описание информационных мер связи можно найти в [Миркин, 1980; Статистические методы ..., 1979; Елисеева, Рукавишников, 1977].

### 2.3.4. Коэффициенты связи для четырехклеточных таблиц сопряженности. Отношения преобладаний

Четырехклеточные таблицы – это частотные таблицы, построенные для двух дихотомических признаков. Встает вопрос – надо ли изучать эти таблицы отдельно? Ведь они представляют собой частный случай всех возможных таблиц сопряженности. Выше мы обсуждали коэффициенты, которые можно использовать для анализа любой частотной таблицы, в том числе и для четырехклеточной. Однако ответ на наш вопрос положителен. Причин тому несколько.

Во-первых, многие известные коэффициенты для четырехклеточных таблиц оказываются равными друг другу. И по крайней мере надо знать об этом, чтобы не осуществлять заведомо ненужные выкладки.

Во-вторых, оказыватся, что именно в анализе четырехклеточных таблиц можно увидеть нечто полезное для социолога, но не высвечивающееся на таблицах большей размерности.

В-третьих, с помощью анализа специальным образом организованных четырехклеточных таблиц оказывается возможным перейти от изучения глобальных связей к изучению локальных и промежуточных между первыми и вторыми (о промежуточных связях мы говорили в п.2.2.1).

Итак, рассмотрим два дихотомических признака – X и Y, принимающие значения 0 и 1 каждый, и отвечающую им четырехклеточную таблицу сопряженности (табл. 14).

Ниже будем использовать пример, когда рассматриваются два дихотомических признака – пол (1 – мужчина, 0 – женщина) и курение (1 – курит, 0 – не курит) (см. табл. 15).

*Таблица 14.*

**Общий вид четырехклеточной таблицы сопряженности**

| X | Y | | Итого |
|---|---|---|---|
| | 1 | 0 | |
| 1 | a | b | a+b |
| 0 | c | d | c+d |
| Итого | a+c | b+d | a+b+c+d |

буквы в клетках обозначают соответствующие частоты

*Таблица 15.*

**Пример четырехклеточной таблицы сопряженности**

| Курение | Пол | | Итого |
|---|---|---|---|
| | м | ж | |
| Курит | 80 | 4 | 84 |
| Не курит | 10 | 6 | 16 |
| Итого | 90 | 10 | 100 |

Данные таблицы 15 говорят о том, что в нашей совокупности имеется 90 мужчин, из которых 80 человек курят, и 10 женщин, среди которых 4 человека курящих и т.д.

Все известные коэффициенты связи для четырехклеточных таблиц основаны на сравнении произведений ad и bc. Если эти произведения близки друг к другу, то полагаем, что связи нет. Если они совсем не похожи – связь есть. Основано такое соображение на том, что равенство $ad = bc$ эквивалентно равенству $\frac{a}{c} = \frac{b}{d}$, что, в свою очередь, означает пропорциональность столбцов (строк) нашей частотной таблицы, т.е отсутствие статистической связи. Чем более отличны друг от друга указанные произведения, тем менее пропорциональны столбцы (строки) и, стало быть, тем больше оснований имеется у нас полагать, что переменные связаны. Для обоснования этого утверждения могут быть использованы те же рассуждения, что были приведены выше. А именно, можно показать, что разница между наблюдаемой и теоретической частотой для левой верхней клетки нашей четырехклеточной частотной таблицы (нетрудно проверить, что наличие или отсутствие связи для такой таблицы определяется содержанием единственной клетки - при заданных маргиналах частоты, стоящие в других клетках, можно определить однозначно) равна величине [Кендалл, Стьюарт, 1973. С. 722]:

$$D = \frac{ad - bc}{n}$$

Коэффициенты, основанные на описанной логике, могут строиться по-разному. Но всегда они базируются либо на оценке разности ($ad - bc$), либо на оценке отношения $\frac{ad}{bc}$. В первом случае об отсутствии связи будет говорить близость разности к нулю, во втором – близость отношения к единице. Естественно, ни разность, ни отношение не могут служить искомыми коэффициентами в “чистом” виде, поскольку их значения зависят от величин используемых частот. Требуется определенная нормировка. И, как мы уже оговаривали выше, желательно, чтобы искомые показатели связи находились либо в интервале от -1 до 1, либо – от 0 до 1, Возможны разные ее варианты. Это обуславливает наличие разных коэффициентов –

показателей связи для четырехклеточных таблиц. Рассмотрим два наиболее популярных коэффициента.

*Коэффициент ассоциации Юла*:

$$Q = \frac{ad - bc}{ad + bc}$$

и *коэффициент контингенции*

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Коротко рассмотрим их основные свойства.

Оба коэффициента изменяются в интервале от -1 до +1 (значит, для них имеет смысл направленность связи; о том, что это такое в данном случае, пойдет речь ниже). Обращаются в нуль в случае отсутствия статистической зависимости, о котором мы говорили выше (независимость признаков связана с пропорциональностью столбцов таблицы сопряженности). А вот в единицу (или - 1) эти коэффициенты обращаются в разных ситуациях. Они схематично отражены ниже.

| Свойства коэффициентов: | Q = 1 | Q = -1 | Φ = 1 | Φ = -1 |
|---|---|---|---|---|
| Отвечающие им виды таблиц | a 0<br>c d<br>a b<br>0 d | 0 b<br>c d<br>a b<br>c 0 | a 0<br>0 d | 0 b<br>c 0 |
| | (а) | (б) | (в) | (г) |

**Рис. 16. Схематическое изображение свойств коэффициентов Q и Φ.**

Таким образом, мы видим, что Q обращается в 1, если хотя бы один элемент главной диагонали частотной таблицы равен 0. Для обращения же в 1 коэффициента Φ необходимо обращение в 0 обоих элементов главной диагонали. Нужны ли социологу оба коэффициента? Покажем, что каждый из них позволяет выделять свои закономерности. Или, как мы говорили выше – за каждым из них стоит своя модель изучаемого явления, свое понимание связи, выделение как бы одной стороны того, что происходит в реальности. Постараемся убедить читателя, что социолога должны интересовать обе эти стороны.

Предположим, что в нашем распоряжении имеется лишь коэффициент Φ и мы даем задание ЭВМ для каких-то массивов данных выдать нам все такие четырехклеточные таблицы,

для которых этот коэффициент близок к единице (может быть, мы хотим найти все те признаки, для которых имеется связь для респондентов некоторой фиксированной совокупности, а, может быть – изучаем, для каких совокупностей респондентов имеется сильная связь между какими-то конкретными признаками). ЭВМ выдаст нам набор таблиц типа (в) или (г). Мы будем знать, к примеру, что имеются группы респондентов, для которых имеется сильная связь между полом и курением: все мужчины курят, а все женщины не курят (что довольно распространено) или наоборот – все женщины курят, а мужчины – нет (что имеет место, скажем, для некоторых индейских племен). Но мы "не заметим", что для каких-то групп все мужчины курят, в то время как среди женщин встречаются и курящие, и не курящие, либо все женщины не курят, хотя мужчины ведут себя по-разному - могут и курить, и не курить (случай (а)). Думается, что не требует особого доказательства утверждение о том, что социолог, не умеющий выискивать подобные ситуации, рискует много потерять. Аналогичное утверждение справедливо и относительно ситуаций, обохзначенных буквой (б).

Другими словами, не используя коэффициент Q, социолог рискует не заметить интересующие его закономерности. Перефразируя сказанное выше вспомнив, что связь также имеет отношение и к прогнозу, отметим, что эти не замеченные закономерности отвечают ситуациям, когда мы по одному значению первого признака можем прогнозировать значение второго, а по другому значению не можем: скажем, зная, что респондент - мужчина, мы с полной уверенностью можем сказать, что он курит, а зная, что респондент - женщина - никакого прогноза, вообще говоря, делать не можем (нижняя таблица случая (а)). Вряд ли можно сомневаться, что выявление и такой "половинчатой" возможности прогноза для социолога может быть полезной.

Рассмотрим теперь вопрос: не можем ли мы обойтись без коэффициента Ф? Представляется очевидным отрицательный ответ на него: выявляя значимые ситуации только с помощью Q, мы можем "за деревьями не увидеть леса" - не заметить, что в отдельных случаях мы может прогнозировать не только по одному значению того или иного признака, но и по другому тоже.

Описанное различие между коэффициентами Q и Ф нашло свое отражение в терминологии. Та связь, которую отражает Q, была названа *полной,* а та, которую отражает Ф, - *абсолютной.*

Еще раз определим эти виды связи, несколько видоизменив формулировку. Для этого вспомним, что, зная маргиналы четырехклеточной таблицы сопряженности, о связи между двумя дихотомическими признаками можно судить по одной частоте. Чаще всего для этого

используют $n_{11}$. Обозначим отвечающие этой частоте значения наших признаков через А и В. Например, А означает “мужчина”, а В – “курит”. В таком случае говорят, что связь между А и В полная, если все А являются одновременно В, несмотря на то, что не все В являются одновременно А (все мужчины курят, но не все курящие являются мужчинами). Если же все А являются одновременно В и все В являются одновременно А (т.е. если все мужчины курят и все курящие – мужчины), то связь называется абсолютной. Иногда для обозначения тех же свойств рассматриваемой связи используют иную терминологию – говорят, что Q измеряет одностороннюю связь, а Ф – двустороннюю.

Поясним теперь, в чем смысл знака рассматриваемой связи. Для этого заметим, что приведенные выше рассуждения можно переформулировать, говоря не о том, что все А являются одновременно В, а о том, что свойства А и В сопрягаются друг с другом (таблица *сопряженности* потому так и названа, что ее придумали для того, чтобы изучать, какие значения разных признаков “ходят” вместе, сопрягаются друг с другом). Термины “положительный” и “отрицательный”, испобзуемые для характеристики связи, носят весьма относительный характер: “положительность” означает, что какое-то значение первого признака сопрягается с одним значением другого, а “отрицательность” – с другим (при наличии положительной связи все мужчины курят, и при наличии отрицательной – все мужчины не курят).

Однако сказанное становится весьма нечетким утверждением при отсутствии нулевых клеток в таблице сопряженности. Например, трудно понять, с каким значением признака "курит – не курит" сопрягается мужской пол, если данные представлены таблицей:

*Таблица 16*

**Частотная таблица для демонстрации отношения преобладаний**

| Курение | Пол | | Итого |
|---|---|---|---|
| | м | ж | |
| Курит | 50 | 90 | 140 |
| Не курит | 20 | 40 | 60 |
| Итого | 70 | 130 | 200 |

С одной стороны, среди курящих больше женщин, чем мужчин. И среди женщин больше курящих, чем некурящих. Но правильно ли будет сказать, что свойство "курит" сопрягается с женским полом? Ведь если среди мужчин курящих в 2,5 раза (50:20) больше курящих, чем некурящих, то среди женщин – лишь в 2,25 раза (90:40). Строгое определение положительной и

отрицательной связи можно дать с помощью введения понятия *отношения преобладаний* [Rudas,1998]:

$$\lambda = \frac{50:20}{90:40}$$

или, в общем случае (обозначения – как в таблице 14):

$$\lambda = \frac{a:c}{b:d}$$

Если отношение преобладания больше единицы, то связь называется *положительной,* если меньше единицы – то *отрицательной.* (Отношение преобладания обобщается на многомерный случай, о чем коротко пойдет речь в п. 2.3.5.).

И еще об одном очень важном моменте необходимо сказать. Если мы, используя обозначения 0 и 1 для значений наших признаков, будем интерпретировать эти обозначения как настоящие числа, то, как нетрудно проверить, вычисленный по обычным правилам коэффициент корреляции между признаками окажется равным Ф. Будучи обобщенным, этот факт имеет огромное значение для анализа данных. Дело в том, что одним из популярных способов создания возможности использования числовых математико-статистических методов для анализа номинальных (нечисловых!) данных является т.н. дихотомизация последних: замена (по определенным правилам) одного номинального признак таким количеством дихотомических, принимающих значения 0 и 1, сколько в этом признаке альтернатив и дальнейшая "работа" с этими 0 и 1 как с обычными числами. Этот подход не имеет строгого математического обоснования. Его "оправдание" состоит в том, что все числовые статистики, рассчитанные по обычным правилам, оказывается возможным разумно проинтерпретировать. Именно пример этого мы и видели выше: коэффициент корреляции, вычсленный для 0 и 1, оказался разумной величиной, совпал с Ф. Вернемся к этому в п. 2.6.3.

О коэффициентах связи для четырехклеточных таблиц можно прочесть в [Интерпретация и анализ ..., 1987. С.29-30; Лакутин, Толстова, 1990, 1992; Паниотто, Максименко, 1982.С.84-93; Рабочая книга ..., 1983. С.189; Статистические методы ... 1979. С.116-117; Libetrau, 1989]

### 2.3.5. Проблема сравнения коэффициентов связи

Заканчивая обсуждение вопроса о коэффициентах связи типа "признак-признак", необходимо упомянуть актуальную для социологии проблему сравнения всех таких коэффициентов. Однако здесь мы не будем ее подробно обсуждать, отнеся читателя к

соответствующей литературе [Елисеева, Рукавишщников, 1982. С.89-101; Интерпретация и анализ ..., 1987.С.34-36; Лакутин, Толстова, 1990, 1992; Миркин, 1980.С.94-109; Паниотто, Максименко, 1982. С.124-125; Рабочая книга ...,1983. С.191-192].

Отметим лишь очень коротко несколько отдельных моментов.

Любой критерий сравнения, как всякий подход к математическому анализу данных, основан на предположениях о том, что реальности адекватны некоторые формальные построения, отражающие определенные аспекты интерпретации исходных данных. Другими словами, для того, чтобы можно было говорить о сравнении, необходимо заранее сформировать некоторую модель того, что мы понимаем под схожими (несхожими) коэффициентами.

Наиболее обоснованное теоретически и часто использующееся в статистической литературе основание для сравнения

226

рассматриваемых коэффициентов базируется на обсужденном выше предположении о том, что за каждым номинальным признаком стоит некоторая латентная непрерывная количественная переменная. Коротко говоря, суть соответствующих подходов заключается в следующем. Исследователь моделирует с помощью ЭВМ некоторую “генеральную совокупность”, описываемую двумя непрерывными переменными с заданным коэффициентом корреляции между ними. Затем упомянутые переменные искусственным образом превращаются в номинальные, из “генеральной” совокупности формируется множество выборок и для каждой из них подсчитываются подлежащие сравнению коэффициенты. Когда выборок организуется достаточно много, появляется возможность сравнения “поведения” отдельных коэффициентов друг с другом.

Сказанное в предыдущих параграфах свидетельствует о том, что все рассмотренные коэффициенты различны. За каждым стоит своя модель, свое понимание этой связи. Вопрос о том, какова же истинная связь между переменными, если такой -то коэффициент равен 0,7, а такой-то - 0,2, не имеет смысла. В описанной ситуации можно сказать только то, что связь в первом смысле (смысле, отвечающем первому коэффициенту) более высока, чем связь во втором смысле. И для того, чтобы найти “истинную” связь, надо использовать целый набор коэффициентов. Каждый их них как бы отвечает отдельной стороне “истины”. А для того, чтобы “истина”, как бриллиант, засверкала всеми своими гранями, необходимо иметь эти грани перед глазами все сразу, “поворачивая” нашу связь в разные стороны.

Однако имеет смысл сказать не только о различии, но и о сходстве разных коэффициентов. Если посмотреть на них с другой стороны, окажется, что не так уж сильно они

расходятся друг с другом. И это не случайно – все-таки речь идет о разных способах формализации одного и того же явления – интуитивно понимаемой связи между переменными. Действительно, можно показать (и это в определенной мере демонстрировалось выше), что так или иначе, в разной степени, но все коэффициенты основаны на представлении о том, что существование связи между двумя

227

признаками означает одновременное соблюдение следующих условий: сильное отклонение от пропорциональности столбцов (строк) исходной таблицы сопряженности; улучшение качества прогнозна значений одного признака при получении информации о значении другого; тот факт, что определенные значения одного признака "любят" встречаться вместе с определенными значениями другого признака. Однако относительно последнего обстоятельства можно заметить следующее (приведем цитату из [Кендалл, Стьюарт, 1973. С. 724]).

"Следует обратить внимание на то, что статистическая связь отличается от связи в обычном смысле. В повседневной речи мы говорим, что A и B связаны, если они достаточно часто встречаются вместе, а в статистике они считаются связанными только в том случае, если A встречается относительно чаще среди B, чем среди не-B. Если 90% курящих страдают плохим пищеварением, то мы не можем сказать, что курение и плохое пищеварение связаны, пока не будет показано, что среди некурящих страдают плохим пищеварением менее, чем 90%." Последнее обстоятельство связано с тем, о чем пойдет речь в следующем параграфе.

### 2.3.6. Учет фактической многомерности реальных связей. Многомерные отношения преобладаний

Коснемся очень важной для практики проблемы, связанной со сравнением коэффициентов не друг с другом, а с некоторыми другими подходами к измерению связи между переменными.

***Актуальность многомерных связей в социологии.***

В реальности двумерных связей практически не существует. Все связи многомерны. Приведем определения.

Связь между тремя переменными называется *трехмерной*, если характер связи между любыми двумя из них зависит от того, каково при этом значение третьей переменной. Связь между четырьмя переменными называется *четырехмерной,* если ее характер для

любых трех признаков зависит от того, каково при этом значение четвертой переменной и т.д. Надеемся ясно, как определяется понятие связи любой размерности.

Многомерность реальных зависимостей заставляет относиться с большой осторожностью к значениями рассмотренных выше парных коэффициентов связи. На это обстоятельство обращают внимание многие исследователи. Поясним это.

В работе [Миркин, 1985. С. 18-20] приводится пример того, как при фиксации значения третьей переменной обусловливает "возникновение" связи между двумя переменными. Опишем его.

Изучалась зависимость между наличием в семьях пылесоса и холодильника. Исходная частотная таблица имела вид:

| | П | ¬П | |
|---|---|---|---|
| Х | 560 | 840 | 1400 |
| ¬Х | 240 | 360 | 600 |
| | 800 | 1200 | 2000 |

Зависимость явно отсутствует, поскольку столбцы (строки) таблицы пропорциональны: $\frac{560}{240}=\frac{840}{360}\frac{1400}{600}=\frac{7}{3}$. Таблицу пересчитали отдельно для двух выделенных среди изучаемой совокупности респондентов групп – для семей с высоким (Д) и низким (¬Д) уровнем дохода. Получились следующие две частотные таблицы:

Для Д

| | П | ¬П | |
|---|---|---|---|
| Х | 520 | 300 | 820 |
| ¬Х | 80 | 100 | 180 |
| | 600 | 400 | 1000 |

Для ¬Д

| | П | ¬П | |
|---|---|---|---|
| Х | 40 | 540 | 580 |
| ¬Х | 160 | 260 | 420 |
| | 200 | 800 | 1000 |

В обоих случаях связь присутствует (пропорциональности строк здесь явно нет). Более того, для первой таблицы она положительна (значение "Х" сопрягается со значением "П": семьи, имеющие холодильник, как правило, имеют и пылесос), а для второй – отрицательна (значение "Х" сопрягается со значением ¬П: семьи, имеющие холодильник, чаще всего не могут купить пылесос).

Вспомнив определение положительной и отрицательной связи через отношение преобладания (п.2.3.4), то же самое выразим более строго. В таблице, отвечающей высокому доходу Д отношение преобладания $\frac{520:80}{300:100}=\frac{13}{6}$ больше единицы, а в таблице, отвечающей низкому доходу аналогичное отношение $\frac{40:160}{540:260}=\frac{13}{108}$ – меньше единицы.

Аналогичный пример, когда статистическая независимость между двумя признаками превращается в зависимость при фиксации значения третьего признака приводится в работе [ДА-система…, 1997. С.181-182].

В [Типология и классификация…, 1982] приводится заимствованный у Лазарсфельда пример того, как фиксация значения третьего признака, напротив, приводит к исчезновению первоначальной двумерной связи.

Речь идет о связи между чтением двух журналов А и Б. Исходная частотная таблица имеет вид (А – респондент читает журнал А, ¬А – не читает, то же для журнала Б):

| | А | ¬А | |
|---|---|---|---|
| Б | 260 | 240 | 500 |
| ¬Б | 140 | 360 | 500 |
| | 400 | 600 | 1000 |

Столбцы не пропорциональны: $\frac{260}{140}=\frac{13}{7}\neq\frac{240}{360}=\frac{2}{3}$

Далее вводится новая переменная – образование респондента (В – высокое, ¬В – низкое). Соответствующие таблицы выглядят так:

Для В

| | А | ¬А | |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Б | 240 | 160 | 400 |
| ¬Б | 60 | 40 | 100 |
| | 300 | 200 | 500 |

Для ¬В

| | А | ¬А | |
|---|---|---|---|
| Б | 20 | 80 | 100 |
| ¬Б | 80 | 320 | 400 |
| | 100 | 400 | 500 |

Нетрудно проверить, что столбцы обеих таблицы пропорциональны, т.е. зависимость в обоих случаях отсутствует. Связь исчезла. В таких случаях говорят, что уровень образования является переменной, объясняющей связь между чтением двух рассматриваемых журналов (здесь мы имеем дело с основным положением, лежащим в основе процесса измерения латентных переменных – с лазарсфельдовской аксиомой локальной независимости; эта аксиома лежит в основе латентно-структурного анализа).

В работе [Аптон, 1982] рассматриваемая проблема обсуждается в исторической ретроспективе. В частности, приводится пример т.н. парадокса Симпсона (1951 год). Приведем соответствующие данные. Исходная таблица имела вид

| | В | ¬В | |
|---|---|---|---|
| А | 495 | 805 | 1300 |
| ¬А | 405 | 295 | 700 |
| | 900 | 1100 | 2000 |

В ней наблюдается явная отрицательная связь: отношение преобладаний $\frac{495:405}{805:295} = 0,45$ – меньше единицы (значение А имеет б**о**льшую тенденцию встречаться с ¬В, чем с В). А в тех двух таблицах, которые получаются в результате фиксирования значения третьего дихотомического признака С оба отношения преобладаний больше единицы, т.е. говорят о положительной связи. Эти таблицы выглядят так:

Для С

| | В | ¬В | |
|---|---|---|---|

| | | | |
|---|---|---|---|
| A | 95 | 800 | 895 |
| ¬A | 5 | 100 | 105 |
| | 100 | 900 | 1000 |

Для ¬ C

| | B | ¬B | |
|---|---|---|---|
| A | 400 | 5 | 405 |
| ¬A | 400 | 195 | 595 |
| | 800 | 200 | 1000 |

Соответствующие же отношения преобладаний равны:

$$\frac{95:5}{800:100} = \frac{19}{8} \qquad \text{и} \qquad \frac{400:400}{5:195} = 39{,}0$$

***Многомерные отношения преобладаний.***

Как это уже неоднократно имело место в наших рассуждениях, все приведенные соотношения в реальности теряют смысл из-за того, что мы имеем дело лишь со статистическими закономерностями. Что значат выражения типа: "при фиксации значения третьей переменной связь между первыми двумя исчезла"? Ведь и при наличии связи отклонение от пропорциональности столбцов носит лишь относительный характер, и при отсутствии связи у нас все же, как правило, пропорциональность не "чистая". Чтобы справиться с неопределенностью, можно использовать отношения преобладаний, введенные нами в п. 2.3.4. Однако требуется их обобщить на многомерный случай. Сделаем это.

Вообще говоря, отношения преобладаний могут быть определены для таблиц любой размерности, в том числе и для одномерных, т.е. для линейных частотных распределений (правда, мы предполагаем, что имеем дело с дихотомическими признаками). Чтобы ввести строгое определение отношения преобладаний, введем новые обозначения.

Сначала предположим, что в нашем распоряжении имеется лишь один признак. Тогда будем обозначать через $P_1$ долю объектов, обладающих первым его значением, а через $P_2$ – вторым. Соответствующее отношение преобладания первого порядка, выражаемое формулой

$$\lambda_1 = \frac{P_1}{P_2},$$

естественно, будет обозначать, во сколько раз объем первого множества больше (меньше) второго. Если отношение преобладания больше 1, говорим о положительном преобладании, если меньше – об отрицательном.

Теперь будем считать, что у нас два дихотомических признака. Через $P_{11}$ обозначим долю объектов с первым значением первого признака и первым значением второго, через $P_{12}$ – с первым значением первого и вторым значением второго и т.д. Двумерная частотная таблица приобретет вид:

| | |
|---|---|
| $P_{11}$ | $P_{12}$ |
| $P_{21}$ | $P_{22}$ |

Легко видеть, что отношение преобладания второго порядка (определенное нами в п.2.3.4 и названное там просто отношением преобладания) конструируется следующим образом.

Фиксируем первое значение второго признака и рассчитываем для соответствующей частотной таблицы отношение преобладания первого порядка:

$$\frac{P_{11}}{P_{21}}$$

То же делаем при фиксации второго значения второго признака:

$$\frac{P_{12}}{P_{22}}$$

Отношением преобладания второго порядка называется отношение первой дроби ко второй.

$$\lambda_2 = \frac{P_{11} : P_{21}}{P_{12} : P_{22}}.$$

Надеемся, смысл его очевиден: мы проверяем, в какой мере столбцы таблицы сопряженности являются пропорциональными. Если $\lambda_2$ равно единице, то двумерной связи нет. Если больше единицы, то говорят о положительной связи (и чем больше отличие от 1, тем больше эта связь). Если $\lambda_2$ меньше 1, то говорят об отрицательной связи.

Итак, $\lambda_2$ – это отношение двух $\lambda_1$ для первого признака – вычисленных отдельно для каждого из двух значений второго признака. Та же логика продолжается дальше. Вводим третий признак с двумя значениями. Фиксируем его первое значение и вычисляем $\lambda_2$ по первым двум признакам (формула та же, что выше выражала $\lambda_2$, но ко всем обозначениям частот добавляется

третий индекс, равный 1; это означает, что все величины отвечают первому значению третьего признака):

$$\frac{P_{111} : P_{211}}{P_{121} : P_{221}}$$

Аналогичную величину вычисляем, фиксируя второе значение третьего признака:

$$\frac{P_{112} : P_{212}}{P_{122} : P_{222}}$$

Находим отношение последних двух величин. Это и будет отношение преобладания третьего порядка:

$$\lambda_3 = \frac{\dfrac{P_{111} : P_{211}}{P_{121} : P_{221}}}{\dfrac{P_{112} : P_{212}}{P_{122} : P_{222}}}$$

Если отношения преобладания второго порядка, вычисленные для каждого из двух значений третьего признака, были примерно одинаковыми, то $\lambda_3$ будет примерно равно 1. Это означает отсутствие трехмерной связи. Если $\lambda_3$ больше 1, говорят о положительной трехмерной связи. Если $\lambda_3$ меньше – об отрицательной трехмерной связи и т.д.

Отношения преобладаний играют огромную роль при анализе номинальных данных. Далее учет многомерности фактически встречающихся в социальной реальности связей становится одной из наших главных задач

## 2.4. Связь типа "альтернатива-альтернатива"

### 2.4.1. Смысл локальной связи. Возможные подходы к ее изучению

Напомним (см. п.2.2.1), что под локальной связью мы понимаем связь между отдельными альтернативами рассматриваемых признаков. Можно ее понимать и более широко. Так, выше, при обсуждении прогнозных и информационных коэффициентов связи мы говорили о том, что знание какого-то одного значения X может нам дать очень большую информацию об Y, а для другого значения X аналогичная информация может быть мала. Это и означает, что для первого значения X имеет место сильная локальная связь.

Сами термины “локальный” и “глобальный” применительно к пониманию связи между переменными, вероятно, впервые были использованы в [Чесноков, 1982].В п. 2.2.1 мы уже

упоминали, что “локальному” подходу в этой работе отвечает понимание связи как некоторого отношения между двумя конкретными градациями a и b признаков X и Y соответственно. В таком случае мы можем говорить о сильной связи, если из того, что для некоторого объекта первый признак принимает значение a, с большой вероятностью следует, что второй признак для того же объекта принимает значение b. И можно говорить о слабой связи, если аналогичная вероятность мала (еще раз напомним, что “глобальная” связь - это результат определенного “усреднения” подобных локальных связей).

Для изучения локальной связи можно использовать, например, коэффициенты Ф и Q. Для этого надо исходную частотную таблицу произвольной размерности привести к определенной четырехклеточной. Покажем на примере, как это делается. Рассмотрим частотную таблицу, выражающую зависимость между

*Таблица 17.*

**Пример таблицы сопряженности**

| Профессия | Читаемая газета | | | | Итого |
|---|---|---|---|---|---|
| | УГ | МК | Независимая | Правда | |
| Врач | 5 | 2 | 13 | 8 | 28 |
| Токарь | 6 | 24 | 7 | 13 | 50 |
| Учитель | 9 | 0 | 1 | 0 | 10 |
| Космонавт | 2 | 1 | 4 | 5 | 12 |
| Итого | 22 | 27 | 25 | 26 | 100 |

профессией человека и читаемой им газетой (для простоты предполагаем, что каждый респондент может читать не более одной газеты). Предположим, что нас интересует локальная связь между свойством “быть учителем” и свойством “читать "Учительскую газету" (УГ)”. Упомянутая выше четырехклеточная таблица будет иметь вид:

*Таблица 18.*

**Четырехклеточная таблица сопряженности, полученная из таблицы 17**

| Профессия | Читаемая газета | | Маргиналы по строкам |
|---|---|---|---|
| | УГ | Не УГ | |
| Учитель | 9 | 1 | 10 |
| Не учитель | 13 | 77 | 90 |

| | | | |
|---|---|---|---|
| Маргиналы по столбцам | 22 | 78 | 100 |

Представляется очевидным, что если мы далее будем использовать коэффициенты связи, предназначенные для анализа четырехклеточных таблиц, то как раз и измерим силу нашей локальной связи.

### 2.4.2. Детерминационный анализ (ДА). Выход за пределы связей рассматриваемого типа

В [Чесноков, 1982] для обозначения того объекта, который является носителем локальной связи, вводится понятие детерминации, обозначаемой $a \to b$ (отметим, однако, что мы несколько вольно трактуем указанное определение, поскольку автор названной работы принципиально отвергает связь детерминации с вероятностью, говоря только об относительных частотах; о них ниже пойдет речь, и мы их будем расценивать как выборочные оценки соответствующих условных вероятностей). Детерминация определяется как носитель локальной связи или как нечто, задаваемой двумя величинами: интенсивностью (точностью, истинностью) $I(a \to b) = P(b / a)$ и емкостью (полнотой) $C(a \to b) = P(a / b)$ (справа стоят относительные частоты).

Рассмотрим приведенную выше таблицу и детерминацию (учитель $\to$ УГ). Интенсивность и емкость в этом случае будут выглядеть следующим образом:

$$I(a \to b) = P(b / a) = P(\text{\it УГ} / \text{\it учитель}) = \frac{9}{10} = 0{,}9$$

$$C(a \to b) = P(a / b) = P(\text{\it учитель} / \text{\it УГ}) = \frac{9}{22} \approx 0{,}41$$

Итак, если мы хотим полностью охарактеризовать связь между свойством “быть учителем” и свойством “читать УГ”, то должны учесть два числа - долю читающих УГсреди учителей (90%) и долю учителей среди читающих УГ (41%). При всей своей простоте, это соображение далеко не всегда учитывается социологами. Частая ошибка применительно к нашему случаю означает, что исследователь узнает, что почти все учителя читают УГ и делает вывод, состоящий в том, что аудитория УГ в основном состоит из учителей. Конечно, логика здесь “хромает” - действительно, учителя составляют менее половины аудитории УГ.

Таким образом, для полного изучения "взаимодействия" двух альтернатив (т.е. изучения детерминации) необходимо принимать во внимание обе величины - и емкость, и интенсивность детерминации. Казалось бы, это достаточно очевидное положение. Тем не менее, социолог часто на практике про это забывает (или хочет "забыть"?!). Приведем пример того, как это обстоятельство приводит к неправильной интерпретации исследователем имеющихся в его распоряжении данных.

После выборов в государственную Думу, прошедших в декабре 1995 года, во многих средствах массовой информации обыгрывался тот факт, что среди голосовавших за КПРФ была относительно мала доля людей с высшим образованием. Действительно, она была меньше, чем аналогичная доля среди голосовавших, скажем за Яблоко или НДР. Естественно, из этого обстоятельства делался вывод о том, что образованные люди не голосуют за компартию.

Но, вспоминая наши показатели, можно сказать, что, делая этот вывод, журналисты опирались только на сравнение величин емкостей детерминаций

(высшее образование) → (голосование за КПРФ),

(высшее образование) → (голосование за "Яблоко"),

(высшее образование) → (голосование за НДР),

т.е. на величины долей людей с высшим образованием среди голосовавших за разные партии. Однако обратимся к анализу интенсивностей тех же детерминаций. Оказывается, что за компартию в декабре проголосовало 1, 54 миллиона избирателей с высшим образованием, за "Яблоко" - 1, 43 миллиона, за НДР - 1, 3 миллиона ("Советская Россия", 21 марта 1996 года). Другими словами, среди лиц с высшим образованием доля проголосовавших за КПРФ (т.е. емкость первой детерминации), больше, чем доля проголосовавших за "Яблоко" и НДР (т.е. емкости второй и третьей детерминации). Так за кого голосуют люди с высшим образованием? Предыдущий вывод вряд ли справедлив.

Вычисление интенсивности и емкости изучаемых детерминаций – основной элемент детерминационного анализа. При всей своей простоте этот подход заключает в себе глубокий смысл, поскольку требование обязательного вычисления названных показателей является своеобразной защитой от недосмотра социологов.

Кроме того, детерминационный анализ не сводится в анализу тех связей, которые мы назвали связями типа “альтернатива-альтернатива”. Он включает в себя целую систему алгоритмов, позволяющих повышать интенсивность и емкость рассматриваемых детерминаций, за счет учета значений множества признаков. Поясним подробнее, о чем здесь идет речь. Однако сначала отметим, что иногда в рамках детерминационного анализа используется

терминология, несколько отличная от приведенной выше: интенсивность детерминации называется ее точностью, емкость – полнотой, сама детерминация – правилом "Если *a*, то b". "*a*" называется при этом объясняющим признаком, "b" – объясняемым. Замети, что Термин “признак” здесь используется в том смысле, который мы придавали словосочетанию “значение (альтернатива, градация) признака”. Надеемся, такое смешение терминов в данном параграфе не приведет к недоразумениям.

Предполагается, что в качестве объясняющего признака могут выступать конъюнкции и дизъюнкции любых значений рассматриваемых признаков-предикторов. При этом совокупность последних является “плавающей”. Все признаки-предикторы в таком случае называются объясняющими.

Процитируем некоторые положения из [Да-система…,1997. С. 160-161].

“Точность правила “Если *a*, то b” вычисляется по формуле:

$$\frac{N(a,b)}{N(a)},$$

где N (*a*,b) – количество объектов, обладающих одновременно объясняющим признаком *a* и объясняемым признаком b (количество подтверждений правила); N(*a*) – количество объектов, обладающих объясняющим признаком *a* безотносительно к любым другим признакам (количество применений правила). Точность измеряется от 0 до 1. Точность правила “Если *a*, то b” есть мера достаточности *a* для наличия b. Точность правила – это главный критерий его практической ценности. Наиболее ценятся правила, имеющие точность, близкую к 1.

Полнота правила – это мера его единственности. Она вычисляется по формуле:

$$\frac{N(a,b)}{N(b)},$$

Где N (b) – количество объектов, обладающих объясняемым признаком b безотносительно к любым другим признакам (объем объясняемого признака). Полнота изменяется от 0 до 1. Полнота правила “Если *a*, то b” есть мера необходимости а для наличия b. Полнота правила – это второй по значимости (после точности) критерий его практической ценности. Предельно точные правила ценятся тем выше, чем больше их полнота. Однако наличие высокой полноты не обязательно. Система точных правил, каждое из которых имеет небольшую полноту, может иметь чрезвычайную полезность для практики и науки, если ее суммарная полнота близка к 1”.

Пакет, реализующий детерминационный анализ [Да-система…,1997], позволяет эффективно подбирать конъюнкции объясняющих признаков для повышения точности правила, дизъюнкции – для повышения его полноты.

Например, предположим, что объясняемое положение – голосование за кандидата N. Допустим, что 40% мужчин проголосовали за N. Это значит, что точность правила “если мужчина, то голосует за N“ равна 0,4. Если мы рассмотрим мужчин с высшим образованием, точность детерминации может повыситься (а может, конечно, и не повыситься, и даже понизиться). Так, например, может оказаться, что за N проголосовали 80% мужчин с высшим образованием. Это будет означать, что, взяв конъюнкцию значения признака “пол”, означающее мужчину, и значения признака “образование”, отвечающее высшему образованию, мы повысили точность детерминации по сравнению с тем случаем, когда не учитывали образование респондента. Аналогичные рассуждения справедливы для полноты детерминации : ее тоже можно повышать с помощью удачного подбора объясняющих признаков.

Для сравнения ДА с другими алгоритмами, решающими сходные задачи, необходимо упомянуть еще два определения из [Да-система…,1997. С.161-162].

“Если какой-либо объясняющий признак убрать из правила, точность правила, вообще говоря, изменится. Величина этого изменения (с учетом знака) и есть, по определению, вклад объясняющего признака в точность. Рассмотрим правило “если a и b, то c". Вклад S (a) объясняющего признака в точность вычисляется по формуле

S(a) = (Точность правила "если *a* и b, то c“) – (точность правила “если b, то c").

Аналогично вычисляется вклад любого объясняющего признака в точность в любом заданном правиле." Совершенно аналогично определяется Вклад Q (a) объясняющего признака в полноту.

Заметим, что пакет программ, реализующий идеи детерминационного анализа на PC (ДА-система), пользуется большоя популярностью у социологов.

Более подробно мы не будем рассматривать ДА. Автор подхода, разработчики соответствующих программ для ЭВМ активно занимаются его пропагандой среди социологов. Однако в определенной мере мы вернемся к обсужденным положениям в п.п. 2.5.4 и 2.5.5, где попытаемся проанализировать ДА с точки зрения возможностей выявления обобщенных взаимодействий и сравнить его с методами поиска логических закономерностей.

Отметим только один факт, очень важный для нас в методологическом аспекте: автор детерминационного анализа развил его дальше, оригинальным образом обобщив положения

аристотелевской силлогистики и построив стройную математическую теорию, отвечающую естественной логике социолога, “невооруженным глазом” анализирующего частотные таблицы [Чесноков, 1985]. Рождение этой теории является ярким примером того, как социологические потребности могут служить толчком для развития новых ветвей математики.

## 2.5. Анализ связей типа "группа альтернатив - группа альтернатив" и примыкающие к нему задачи

### 2.5.1. Классификация задач рассматриваемого типа

Итак, мы проанализировали суть связей типа "альтернатива × альтернатива", убедились в важности их изучения. Нетрудно видеть, что логика, сходная с использованной выше, приводит к мысли о необходимости изучения подобных связей для таких ситуаций, когда вместо отдельных альтернатив фигурируют их группы. Например, вместо задачи изучения связи между свойствами "быть учителем" и "читать Учительскую газету" мы можем поставить задачу проанализировать зависимость между свойствами "быть учителем, или врачом, или научным сотрудником, или иметь одну из т. н. творческих профессий" и "читать Литературную газету или журнал Новый Мир". Казалось бы, никаких проблем при решении такой задачи не должно возникать. Нужно только рассмотреть отвечающую нашим альтернативам подтаблицу исходной "большой" таблицы сопряженности и применить к ней уже знакомые нам способы измерения связей между двумя номинальными признаками.

Проблемы возникают в том случае, если мы не фиксируем заранее указанную подтаблицу, а ставим перед собой цель, например, найти такие подтаблицы исходной таблицы сопряженности, которые обладают свойствами, отличающими их от всей таблицы (либо от других подтаблиц). Например, такие, для которых тот или иной коэффициент связи больше (меньше), чем на всей таблице (на других подтаблицах). В качестве еще одной цели может служить изучение того, за счет каких подсвязей формируется наша "большая" связь. Можно считать целью изучение каких-то свойств, скажем, не учителей и врачей вместе (т.е. не такого множества респондентов, которое отвечает совокупности значений одного и того же признака - в данном случае - профессии), а, например, учителей старше 50 лет, работающих в гимназиях (т.е. совокупности респондентов, отвечающей набору значений разных признаков - в данном случае - профессии, места работы и возраста). Возможны и другие повороты. Рассмотрим два класса методов, определяемых выбором цели.

*Первый класс методов - группа альтернатив отвечает одному признаку.*

Рассматриваемый класс определяется тем, что каждая из "групп альтернатив", означенных в названии нашего параграфа, состоит из значений одного признака (скажем, это разные наименования профессий, т.е. разные значения признака "профессия"). Исходная информация в таком случае представляет собой таблицу сопряженности между двумя признаками, отвечающими нашим двум "группам альтернатив".

Здесь можно было бы, в свою очередь, говорить о возможности выделения двух подклассов задач.

*Первый подкласс* – математико-статистический. Речь идет о выяснении того, из каких компонент состоит величина "Хи-квадрат", вычисленная для рассматриваемой частотной таблицы, или, как мы будем говорить, о разложении этой величины на составные части, позволяющие определить, какой вклад в нее осуществляют разные фрагменты таблицы сопряженности. Для решения соответствующих задач существуют строгие правила перенесения результатов с выборки на генеральную совокупность и т.д. Этот подкласс будет подробно рассмотрен нами в следующем параграфе.

*Второй подкласс* состоит из типичных задач анализа данных. Для них не разработан тот "антураж", которого требуют строгие каноны математической статистики. Об этом подклассе скажем несколько слов здесь.

Будем полагать, что нас не интересует разложение $\chi^2$, т.е. не интересует выяснение того, из чего состоит эта величина, каков вклад в нее тех или иных фрагментов таблицы сопряженности. Зададимся более простой целью: поиском в этой таблице таких ее подтаблиц, которые отличаются наиболее сильной связью (понимаемой в каком-нибудь из известных нам смыслов) между определяющими эти подтаблицы группами альтернатив. Ясно, что решение этой задачи сводится к простому перебору всевозможных подтаблиц и вычислению отвечающих им показателей связи. Большой науки для этого не требуется. Мы не будем больше рассматривать эту задачу (и, стало быть второй подкласс методов), отметив, однако, ее важность для социолога.

*Второй класс методов – группа альтернатив отвечает разным признакам.*

Методы этого класса также относятся к типичным методам анализа данных, поскольку для них не разработан строгий математико-статистический подход. О них пойдет речь в п. 2.5.3. Мы увидим, что приведенное в заглавии п.2.5 название типа изучаемых связей естественным образом может быть обобщено: во многих реальных ситуациях вместо задач типа "(группа

альтернатив)-(группа альтернатив)" имеет смысл рассмотреть задачи типа "(группа альтернатив)-("поведение" респондентов)", где "поведение" может быть описано не только путем задания отвечающих рассматриваемым респондентам групп альтернатив, но и другими способами.

### 2.5.2. Анализ фрагментов таблицы сопряженности.

Первая задача, которую мы рассмотрим, состоит в своего рода "анатомировании" величины статистики $\chi^2$, вычисленной для нашей исходной таблицы (будем такую статистику называть "большим" $\chi^2$). Попытаемся разложить эту статистику на части, отвечающие каким-то подтаблицам исходной таблицы сопряженности, и понять, какая из этих подтаблиц вносит наибольший вклад в общий $\chi^2$ . Математическая статистика дает нам возможность это сделать.

(Надо сказать, что математика предлагает бесконечное количество различных разложений Хи-квадрата. И отдельные элементы этих разложений совсем не обязательно отвечают каким-то подтаблицам исходной таблицы сопряженности. Разложение может строиться совсем по другому принципу. Но в любом случае за каждым членом разложения стоит какой-то определенный аспект, срез некоторого общего понятия связи. Здесь мы не имеем возможности объяснить это более подробно. Отметим лишь то, что в более полном курсе мы рассматриваем метод канонического анализа таблиц сопряженности, который, в частности, включает в себя разложение $\chi^2$, не отвечающее разбиению исходной таблицы на части.)

Существует возможность такого разложения исходной частотной таблицы на четырехклеточные подтаблицы, что исходный "большой" Хи-квадрат будет приблизительно равен сумме "четырехклеточных" Хи-квадратов. При этом количество упомянутых подтаблиц равно числу степеней свободы исходной таблицы. Другими словами, при использовании рассматриваемого подхода будет иметь место приблизительное равенство

$$\chi^2 \approx \sum_i \chi_i^2 \qquad (5)$$

где $\chi_i$ отвечает i-й четырехклеточной компонентной подтаблице (т.е. подтаблице, являющейся одной из компонент разложения исходной таблицы сопряженности). Чтобы понять смысл такого разложения, вспомним, что величина Хи-квадрат есть величина отклонения теоретических частот (т.е. тех, которые должны были бы иметь место при условии статистической независимости рассматриваемых признаков, при пропорциональности столбцов (строк) таблицы сопряженности) от эмпирических. При расчете этого показателя мы как бы

суммируем, усредняем отдельные "клеточные" отклонения. А ведь они могут быть разными: в одних клетках наблюдаемые частоты могут совпадать с теоретическими, в других - сильно от них отличаться. Соответственным образом могут отличаться друг от друга не только отдельные клетки, но и другие фрагменты исходной таблицы сопряженности. В интересующем нас случае рассматриваются не произвольные фрагменты, а лишь четырехклеточные. И соотношение (5) говорит о том, какой именно вклад в общее отклонение частот от условия статистической независимости дают фрагменты такого рода.

Что же практически нам дает разложение (5)? Ничего, если все "четырехклеточные" Хи-квадраты превышают (или все – не превышают) соответствующие табличные критические значения (т.е. если для всех наших компонентных подтаблиц мы должны отвергнуть (или для всех же – принять) нуль-гипотезу о независимости соответствующих пар альтернатив друг от друга. Очевидно, что в таком случае и исходный "большой" Хи-квадрат превышает (не превышает) отвечающее ему табличное значение (напомним, что подобные критические значения будут разными у исходной таблицы и у рассматриваемых компонентных подтаблиц, поскольку они имеют разное число степеней свободы) и мы можем считать, что отвержение (принятие) соответствующей нуль-гипотезы как бы равномерно опирается на все значения рассматриваемых признаков. Считаем, что в таком случае никаких интересующих нас подсвязей исходная таблица сопряженности не содержит.

Другое дело, если одни "четырехклеточные" Хи-квадраты будут превышать соответствующие критические значения, а другие – не будут. Скажем, если окажется, что из десяти полученных компонентных подтаблиц только для трех имеются основания отвергнуть отвечающую им нуль-гипотезу, то это будет означать, что наш исходный "большой" Хи-квадрат отличается от нуля (показывает отклонение ситуации от состояния статистической независимости признаков) за счет наличия связи именно в этих трех подтаблицах, остальные же подтаблицы к наличию связи не имеют отношения.

Прежде, чем привести конкретный пример того, какую прибавку к нашим знаниям о взаимосвязях изучаемых признаков может дать использование рассматриваемого подхода, коротко опишем, каким образом должно строиться интересующее нас разложение исходной таблицы сопряженности. Но сначала отметим, что термин "подтаблица" в данном случае понимается своеобразно. А именно, подтаблица может получаться не только за счет буквального "вырезания" соответствующего фрагмента из исходной матрицы сопряженности, но и в результате суммирования определенных строк и столбцов последней. Примером может служить то, как выше мы для изучения связи свойств "быть учителем" и "читать Учительскую

газету" получали из исходной таблицы (табл. 16) четырехклеточную таблицу сопряженности (табл. 17): в клетке, отвечающей сочетанию "не учитель, читает УГ" стояла частота, полученная из исходной таблицы путем суммирования всех респондентов, читающих УГ, но имеющих профессии, отличные от профессии учителя и т.д. Схематично соответствующую таблицу можно изобразить так:

*Таблица 19.*

**Схематическое изображение четырехклеточного фрагмента таблицы 17**

| | Читает УГ | Не читает УГ |
|---|---|---|
| Учитель | Исходная частота | Сумма респондентов-учителей, читающих газеты, отличные от УГ |
| Не учитель | Сумма респондентов, являющихся не учителями и читающих УГ | Сумма респондентов, являющихся не учителями и читающих газеты, отличные от УГ |

Учитывая это, а также вспоминая, что понятие маргинальной суммы имеет смысл не только для исходной таблицы, но и для всех ее подтаблиц, сформулируем правила получения интересующих нас ее компонентных четырехклеточных фрагментов (эти правила мы заимствуем у И. И. Елисеевой [Интерпретация и анализ …., 1987, с.43-44]).

1. Каждая из частот исходной таблицы должна встречаться только в одной из компонентных таблиц.

2. Маргинальные частоты исходной таблицы должны встречаться в одной из компонентных таблиц как частоты определенного типа: либо как "клеточные" (т.е. стоящие в клетке частотной таблицы), либо как маргинальные.

3. Каждая частота, содержащаяся в одной из компонентных таблиц, но отсутствующая в исходной таблице (а такие могут встретиться в тех специфических подтаблицах, о которых мы говорили выше) должна появится в другой компонентной таблице как частота другого типа: "клеточная", если была маргинальной, и наоборот.

Отметим, что сформулированные правила не определяют разложение однозначным образом. То, какое из возможных разложений мы выберем для интерпретации, определяется содержательными соображениями. Возможна и такая ситуации, когда мы усмотрим нечто содержательно полезное в нескольких разложениях. Перейдем к примеру. Воспользуемся цитированной выше работой.

Итак, следуя И. И. Елисеевой, рассмотрим задачу изучения по данным обследования семейных групп (семья сына или дочери - семья родителей) зависимости характера желаемого

расселения (отделения "молодой" семьи от семьи родителей) от состава "молодой" семьи и возраста женщины в этой семье. Исходная частотная таблица имеет следующий вид:

*Таблица 20.*

**Таблица сопряженности, используемая для разложения ее на четырехклеточные подтаблицы**

| Характеристика "молодой" семьи | | Желаемое расселение | | | Итого |
|---|---|---|---|---|---|
| Возраст женщины (лет) | состав | в одной квартире | в разных квартирах | в одном микр-не и дальше | |
| До 30 | Мать с детьми | 6 | 8 | 6 | 20 |
| | Брачная пара с детьми | 11 | 112 | 66 | 189 |
| 30-40 | Мать с детьми | 6 | 12 | 18 | 36 |
| | Брачная пара с детьми | 24 | 122 | 121 | 267 |
| 40-55 | Мать с детьми | 5 | 5 | 8 | 18 |
| | Брачная пара с детьми | 8 | 23 | 8 | 39 |
| Итого | | 60 | 282 | 227 | 569 |

Отметим, что здесь два признака, характеризующие "молодую" семью (ее состав и возраст женщины) фактически превращены в один новый признак, значениями которого служат сочетания значений первоначальных признаков. Именно это позволило таблицу, фактически являющуюся трехмерной, превратить в двумерную. Нетрудно проверить, что на основе вычисления для этой статистики величины $\chi^2$ на 5-процентном уровне значимости можно сделать вывод о том, что у нас имеются все основания отвергнуть нуль-гипотезу об отсутствии статистической связи между нашими двумя признаками: $\chi^2$=39,2, в то время, как $\chi^2_{табл}$ = 18,3 ($\alpha$=0,05; df=10). Встает вопрос: все ли значения рассматриваемых признаков играют одинаковую роль в процессе возникновения этой связи (точнее, в том, что эмпирические частоты оказались отличными от теоретических)? Может ли быть так, что между какими-то наборами альтернатив связь существует, а между какими-то – нет? Чтобы понять это, воспользуемся одним из возможных разложений нашей исходной таблицы на четырехклеточные (в цитируемой нами работе представлено три варианта такого разложения; каждое из них позволяет сделать свои содержательные выводы; мы воспользуемся только тем разложением, которое в названной работе приведено первым).

Для того, чтобы было ясно, как строится разложение (как выделяются четырехклеточные подтаблицы) приведем примеры нескольких таких подтаблиц.

**Разложение таблицы 20 на подтаблицы**

| | | |
|---|---|---|
| 6 | 14 | 20 |
| 54 | 495 | 549 |
| 60 | 509 | 569 |

(А)

| | | |
|---|---|---|
| 8 | 6 | 14 |
| 274 | 221 | 495 |
| 282 | 227 | 509 |

(Б)

| | | |
|---|---|---|
| 11 | 178 | 189 |
| 43 | 317 | 360 |
| 54 | 495 | 549 |

(В)

| | | |
|---|---|---|
| 112 | 66 | 178 |
| 162 | 155 | 317 |
| 274 | 221 | 495 |

(Г)

| | | |
|---|---|---|
| 6 | 30 | 36 |
| 37 | 287 | 324 |
| 43 | 317 | 360 |

(Д)

| | | |
|---|---|---|
| 12 | 18 | 30 |
| 150 | 137 | 287 |
| 162 | 155 | 317 |

(Е)

| | | |
|---|---|---|
| 24 | 243 | 267 |
| 13 | 44 | 57 |
| 37 | 287 | 324 |

(Ж)

| | | |
|---|---|---|
| 122 | 121 | 243 |
| 28 | 16 | 44 |
| 150 | 137 | 287 |

(З)

| | | |
|---|---|---|
| 5 | 13 | 18 |
| 8 | 31 | 39 |
| 13 | 44 | 57 |

(И)

| | | |
|---|---|---|
| 5 | 8 | 13 |
| 23 | 8 | 31 |
| 28 | 16 | 44 |

(К)

Надеемся, читатель сам проследит, какие закономерности лежат в основе формирования приведенных подтаблиц и как в процессе такого формирования реализуются сформулированные выше правила. Перейдем к содержательному анализу подтаблиц.

Не будем приводить разобранный в цитируемой работе пример полностью. Воспользуемся лишь исходной таблицей и двумя полученными при ее разложении подтаблицами. Покажем, какую прибавку к нашим знаниям об изучаемом явлении дает нам рассмотрение этих подтаблиц. При этом мы обратим внимание читателя на такие аспекты упомянутого явления, которые в цитируемой работе не рассматриваются.

Прежде всего отметим, что лишь для 5-ти из 10-ти получившихся четырехклеточных таблиц соответствующее значение $\chi^2$ превышает табличное, отвечающее тому же 5%-му уровню значимости (это значение будет отличаться от приведенного выше из-за различия числа соответствующих степеней свободы: для исходной таблицы это число равно 10, а для четырехклеточной – 1), и равное в данном случае $\chi^2_{табл}$ = 3,8. Чтобы понять, что в содержательном плане может нам дать указанный факт, более подробно опишем

рассматриваемое разложение исходной таблицы сопряженности. Компонентные четырехклеточные таблицы определяются следующими значениями наших признаков:

Заметим, что везде предполагается, что в семье имеются дети, мы же пишем для сокращения "брачная пара" вместо "брачная пара с детьми".

Надеемся, читателю понятно, что частоты, отвечающие значению первого признака "остальные" из таблицы (А), получаются путем суммирования строк исходной таблицы, соответствующих всем рассматриваемым сочетаниям значений двух наших характеристик "молодой" семьи, кроме сочетания "женщина с детьми, до 30 лет"; частоты, отвечающие значению второго признака "в разных квартирах", получаются за счет суммирования столбцов исходной матрицы, отвечающих значениям "в одном доме" и "в одном микрорайоне и дальше" и т.д.

Критический уровень превышают критерии $\chi^2$, отвечающие таблицам (А), (В), (Г), (Ж), (К). Сумма этих критериев равна 33, 9, что, хотя и не равно значению $\chi^2$ для исходной таблицы (напомним, что это значение равно 39, 2), но, как нетрудно проверить, составляет от него почти 86%. Другими словами, отклонение эмпирических частот от теоретических в исходной таблице почти

*Таблица 21.*

**Описание компонентных подтаблиц таблицы 20**

| 1-й признак | 2-й признак | Обозначение подтаблицы |
|---|---|---|
| (мать с детьми, до 30 лет)<br>остальные | в одной квартире,<br>в разных квартирах | (А) |
| то же | в одном доме,<br>дальше | (Б) |
| (брачная пара, мать до 30 лет )<br>остальные | в одной квартире,<br>в разных квартирах | (В) |
| то же | в одном доме,<br>дальше | (Г) |
| (мать с детьми, 30-40 лет)<br>остальные | в одной квартире,<br>в разных квартирах | (Д) |
| то же | в одном доме,<br>дальше | (Е) |

| | | |
|---|---|---|
| (брачная пара, мать 30-40 лет)<br>остальные | в одной квартире,<br>в разных квартирах | (Ж) |
| то же | в одном доме,<br>дальше | (З) |
| (мать с детьми, 40-55 лет)<br>(брачная пара, 40-55 лет) | в одной квартире,<br>в разных квартирах | (И) |
| то же | в одном доме,<br>дальше | (К) |

на 86% объясняется наличием связи в перечисленных четырехклеточных таблицах. Попытаемся на примере показать некоторые "содержательные" аспекты этого положения (чего не было сделано в цитируемой нами работе).

Рассмотрим таблицу (А) (табл. 22).

*Таблица 22.*

**Пример (А) компонентной подтаблицы таблицы 20**

| Тип молодой семьи | Желаемое расселение | | Итого |
|---|---|---|---|
| | В одной квартире | В разных квартирах | |
| Мать с детьми, до 30 лет<br>Остальные | 6<br>54 | 14<br>495 | 20<br>549 |
| Итого | 60 | 509 | 569 |

Значение $\chi^2$ для этой таблицы равно 8,3, что превышает табличное значение, равное 3,8. Нетрудно видеть, что отступление от ситуации независимости (в данном случае мы отождествим ее с пропорциональностью строк) происходит за счет того, что доля желающих остаться в одной квартире со старшим поколением молодых матерей-одиночек (таких молодых матерей-одиночек почти треть: 6 из 20) выше, чем аналогичная доля среди всех опрошенных (среди всех опрошенных не хотят разъезжаться с бабушками-дедушками лишь чуть более 10% : 60 из 569). Вывод – для семей, состоящих из молодых матерей одиночек с детьми, вопрос о необходимости разъезжаться со старшим поколением стоит менее остро, чем для других категорий семей.

Более глубоко можно проанализировать ситуацию с помощью рассмотрения других компонентных таблиц. Ограничимся кратким анализом лишь двух из них: (Б) и (Д) - таких, для

которых соответствующие значения $\chi^2$ (равные, соответственно, 0,02 и 0,8), не превышают критических (см. таблицы 23 и 24).

*Таблица 23.*

**Пример (Б) компонентной подтаблицы таблицы 20.**

| Тип молодой семьи | Желаемое расселение | | Итого |
|---|---|---|---|
| | в одном доме | дальше | |
| Мать с детьми, до 30 лет | 8 | 6 | 14 |
| Остальные | 274 | 221 | 495 |
| Итого | 282 | 227 | 509 |

*Таблица 24.*

**Пример (Д) компонентной подтаблицы таблицы 20.**

| Тип молодой семьи | Желаемое расселение | | Итого |
|---|---|---|---|
| | в одном доме | дальше | |
| Мать с детьми, до 30 лет | 6 | 30 | 36 |
| Остальные | 37 | 287 | 324 |
| Итого | 43 | 317 | 360 |

Для получения интересующих нас выводов достаточно вспомнить, что сравнительно малые значения упомянутого критерия говорят о том, что мы можем считать пропорциональными

252

столбцы (строки), в том числе маргинальные, наших четырехклеточных таблиц. Таблица (Б) (см. табл. 23) говорит о том, что молодые матери-одиночки примерно в той же мере выбирают те или иные варианты расселения, что и семьи других типов. Другими словами соответствующая специфика семьи не сказывается в том, хочет ли желающая переселиться "молодая" семья (нетрудно видеть, что только такие семьи здесь рассматриваются, поскольку во втором признаке задействованы лишь две категории, относящиеся к ситуации разъезда), после переезда остаться поближе к родителям (в одном доме) или же готова уехать подальше. И среди

всех желающих разъехаться чуть более половины хочет остаться в одном доме со старшими (282 из 509), и среди матерей-одиночек до 30 лет (8 из 14).

При анализе таблицы (Д) (см. таблицу 24) становится ясно, что для более старших матерей одиночек – 30-40 лет – указанной выше специфики в желании расселиться нет: семьи этой категории ровно в той же мере хотят разъезда (6 из 36 семей не хотят отделяться от старших), как и семьи других типов (не хотят разъезжаться 37 из 324).

Рекомендуем читателю связать приведенные рассуждения, касающиеся анализа подтаблиц (табл. 21) с анализом соответствующих отношений преобладаний (п.2.3.4).

В заключение параграфа упомянем еще один метод, позволяющий иным путем решать сходные задачи [Ростовцев, 1996, 1998]. Метод предназначен для быстрого обнаружения основных тенденций связи пары переменных. Исходными данными служит совокупность объектов, описанных двумя переменными. В отличие от задачи, рассмотренной выше, здесь предполагается, что используемые шкалы могут быть любыми (в том числе и номинальными). Метод состоит в поиске такой пары дихотомических разбиений совокупностей значений исходных переменных (в результате такого разбиения каждая переменная превращается в дихотомическую), чтобы получающаяся четырехклеточная таблица сопряженности была бы максимально “контрастной”, т.е. отвечала бы как можно более сильной связи между полученными дихотомическими переменными (черно-белый анализ связей).

253

Преимущества подхода ясны – в случае использования метода, описанного выше, мы не имеем гарантий того, что нашли именно те четырехклеточные таблицы, которые характеризуют наиболее сильные дихотомические связи. Здесь же метод позволяет сразу найти именно ту четырехклеточную подтаблицу, которая отвечает максимальной зависимости между конструируемыми дихотомическими переменными. Однако есть здесь и свой минус - мы не можем интерпретировать значение соответствующего (“четырехклеточного”) показателя связи как вклад в величину “большого” критерия, характеризующего связь между исходными перменными. Приведем пример из названной работы, демонстрирующий возможности рассматриваемого подхода.

Рассматривается две переменных: профессиональная подготовка и доходы. Каждой переменной отвечает вопрос в анкете с определенным набором ответов (число которых существенно больше двух; мы сознательно не перечисляем конкретные варианты ответа; они носят довольно стандартный характер и их точная формулировка не является приниципиальной для целей нашего изложения). Проверяется гипотеза о том, что люди, имеющие более высокое

образование, имеют шанс получать более высокие доходы. Автор решил обосновать свою гипотезу путем оценки связи для четырехклеточной таблицы со значениями признаков: высокий доход – низкий доход, высокая профессиональная подготовка – низкая профессиональная подготовка.

Подчеркнем, что стремление свести изучение связи к анализу частотной таблицы минимального возможного размера – четырехклеточной – не является случайным. Напомним читателю, что, во-первых, выявление любой закономерности связано с потерей информации и, во-вторых, сам термин “закономерность” мы применяем только к сравнительно простым, малоразмерным соотношениям.

В рассматриваемой задаче встает вопрос о том, где граница между высоким и низким доходом, между высокой и низкой профессиональной подготовкой. Чаще всего исследователь определяет эту границу интуитивно. Именно это и попытался сделать сначала автор цитируемой статьи. В качестве границы для

254

душевого дохода он взял его среднее значение для изучаемой совокупности респондентов. Уровни профессиональной подготовки были сгруппированы неким естественным образом, при этом ответ “другое” не учитывался. Для проверки своей гипотезы автор получил следующую частотную таблицу:

*Таблица 25.*

**Четырехклеточная таблица, получающаяся в результате “естественного” деления диапазона изменения каждого признака на две части.**

| Душевой доход | Профессиональная подготовка | | Итого |
|---|---|---|---|
| | Невысокая | Высокая | |
| Ниже среднего (менее 5300) | 465<br>81,3% | 276<br>57,1% | 741 |
| Выше среднего (не менее 5300) | 107<br>18,7% | 207<br>42,9% | 314 |
| Итого | 572 | 483 | 1055 |

Проценты означают доли соответствующих совокупностей лиц среди людей с данным уровнем профессиональной подготовки. Нетрудно видеть, что гипотеза подтвердилась: среди

лиц с невысоким уровнем профессиональной подготовки 81,3% людей имеют доход ниже среднего, а среди лиц с высоким уровнем образования – аналогичная доля меньше, 57,1% и т.д.

В качестве критерия оценки степени зависимости душевого дохода респондента от уровня его профессиональной подготовки автор предложил использовать различие между эмпирической и теоретической частотами, отвечающими левой верхней клетке получившейся четырехклеточной таблицы сопряженности. В данном случае критерий равен

$$n_{11}^{э} - n_{11}^{T} = 465 - 401{,}8 = 63{,}2$$

Возник вопрос – нельзя ли подобрать группировку значений переменных, еще ярче подчеркивающую найденную зависимость? И с помощью предложенного в названной статье алгоритма такую группировку удалось найти (табл. 26).

*Таблица 26.*

**Четырехклеточная таблица, получающаяся в результате деления диапазона изменения каждого признака на две части с помощью рассматриваемого алгоритма**

| Душевой доход | Профессиональная подготовка | | Итого |
|---|---|---|---|
| | Невысокая | Высокая | |
| Низкий (менее 4500) | 405<br>71,7% | 203<br>41,4% | 608 |
| Высокий (не менее 4500) | 160<br>28,3% | 287<br>58,6% | 447 |
| Итого | 565 | 490 | 1055 |

Нетрудно проверить, что проверяемая гипотеза подтвердилась более ярко. Это проявилось в том, что здесь оказалось более

255

высоким значение нашего критерия: $n_{11}^{э} - n_{11}^{T} = 405 - 325{,}6 = 79{,}4$. Причина – более удачная группировка людей по доходу.

Заметим, что в ИЭиОПП СО РАН под руководством П.С. Ростовцева разработан пакет программ, реализующий обсужденный подход.

Перейдем к рассмотрению другой ситуации – когда наши группы альтернатив составляются из значений разных признаков. Как мы отмечали, эта ситуация не имеет статистической базы, подобной той, на которую опирается метод анализа фрагментов таблицы сопряженности.

### 2.5.3. Методы поиска сочетаний значений независимых признаков (предикторов), детерминирующих "поведение" респондентов

#### 2.5.3.1. Понятие зависимой и независимых переменных. Общая постановка задачи.

Итак, перед нами огромный массив информации, скажем 1000 заполненных анкет (в таком случае изучаемые объекты – респонденты) по 30 вопросов в каждой (каждому вопросу отвечает признак, описывающий изучаемые объекты). При изучении причинно-следственных отношений естественно выделение, с одной стороны, некоторых признаков, которые описывают

256

основное интересующее исследователя явление, а, с другой – совокупности признаков, потенциально являющихся причинами (напомним, что термин “причина” для нас имеет лишь статистический смысл), обусловливающими то, упомянутое явление имеет именно наблюдаемый вид. Для обозначения признаков первого набора мы по традиции будем использовать букву Y с индексами, а для обозначения признаков второго набора – букву X с индексами. X – независимые переменные (объясняющие, детерминирующие, признаки-причины, аргументы, предикторы), Y – зависимые переменные (объясняемые, детерминируемые, целевые, критериальные, результирующие, признаки-следствия, функции). К этой терминологии мы вернемся в п.2.6. Сейчас же рассмотрим следующую задачу.

Социолога интересует, чем, какими факторами (причинами) определяется некоторое “поведение” респондента. Это “поведение” описывается какими-то признаками Y. Например, оно может состоять в том, что респондент в ответе на один из вопросов анкеты выражает свою готовность проголосовать на выборах за кандидата Ж. Задача состоит в определении того, какими характеристиками (поскольку наша информация о респондентах ограничивается анкетными данными, то этими характеристиками могут быть лишь ответы респондентов на вопросы анкеты) можно описать людей, обладающих рассматриваемым “поведением”, т.е. желающих проголосовать за Ж. Другими словами, мы должны установить какими сочетаниями значений рассматриваемых признаков обладают эти люди.

В принципиальном плане такая задача решается как-будто просто: мы должны перебрать все возможные сочетания значений рассматриваемых признаков и найти среди них такие, обладателям которых присуще рассматриваемое поведение. Схематически это решение можно изобразить следу ющим образом.

Приведенные на схеме стрелки могут означать, к примеру (при соответствующей расшифровке вариантов ответов на вопросы анкеты), что искомым поведением обладают женщины со средним или среднеспециальным образованием, замужние, из семей крестьян или служащих.

257




**Рис. 17. Схематическое изображение сути задач поиска взаимодействий**

Однако в действительности все обстоит не так просто.

Во-первых, перебор всех мыслимых сочетаний значений рассматриваемых признаков столь объемен, что оказывается не под силу даже современным ЭВМ (мы не знаем ни какие признаки взять, ни сколько таких признаков должно быть, ни то, какие сочетания значений каждого признака следует принять во внимание). Встает вопрос о создании определенного алгоритма “сокращенного” перебора. Отметим, что такой алгоритм будет заведомо пропускать определенные сочетания наших независимых признаков; то, какие именно – определяется сутью алгоритма, заложенной в нем моделью, в данном случае связанной с пониманием “поведения” объектов. И для социолога очень важен анализ тех аспектов формализма, которые непосредственно связаны с содержательными аспектами задачи.

Здесь необходимо отметить следующее обстоятельство. Говоря о поведении, мы прежде всего имеем в виду определенное свойство отдельного человека - скажем, то, голосует он или не голосует за того или иного кандидата. Однако в рассматриваемой задаче нам необходимо определить, что такое групповое “поведение”. Ясно, что группу, где 100% людей обладает тем или иным интересующим нас свойством, мы вряд ли найдем из-за принципиальной ненадежности нашего способа измерения мнений респондентов (таким способом для нас является анкетный опрос). Встает вопрос

258

о том, в какой ситуации, рассматривая, скажем, упомянутую выше группу женщин, мы будем иметь право сказать, что нашли совокупность людей с искомым “поведением”. Для используемого примера, вероятно, такую ситуацию естественно связывать с тем, что среди рассматриваемых женщин достаточно высока доля желающих голосовать за Ж. На этом пока и остановимся. Позже вернемся к обсуждению вопроса о других возможных подходах к пониманию группового “поведения”.

Будем называть ту или иную группу респондентов типом, “олицетворяющим” интересующее нас “поведение”, или просто типом, если для этой группы удовлетворяется выбранный нами критерий. Нетрудно видеть, что в случае указанного выше понимая группового поведения мы можем ввести также оценку “качества” группы с точки зрения возможности ее рассмотрения как типа: более высокое качество будет иметь та группа, где доля желающих голосовать за Ж выше. Будем считать, что такая возможность имеется всегда.

Предположим, что упомянутый выше алгоритм сокращенного перебора создан. Тогда "лобовой" путь решения интересующих нас задач будет состоять в следующем: в соответствии с упомянутым алгоритмом перебираются всевозможные сочетания значений рассматриваемых признаков и для каждого из них проверяется, можно ли соответствующую совокупность объектов считать "олицетворением" определенного типа поведения. Если нет – переходим к "проверке" следующего сочетания значений аргументов, если да - считаем, что нашли решение задачи (таких решений может быть много) и в таких случаях группу будем называть типом. Но тут встает еще один вопрос, наше “во-вторых”.

Итак, во-вторых, неясно, как понимать “поведение” группы респондентов. Так, даже для такого простого случая, о котором шла речь выше, неясно, при каких условиях считать, что мы нашли группу, обладающую указанным поведением: если среди этих людей 90% желают проголосовать за Ж? Или 85?

Таким образом, можно сказать, что задача сводится к поиску взаимодействий (определение этого термина дано в п.2.2.1) – сочетаний значений независимых признаков (эти значения, вообще

259

говоря, могут "надергиваться" из разных признаков-предикторов, это – одно из отличий рассматриваемого подхода от подходов, проанализированных в предыдущих параграфах), детерминирующих определенным образом заданное поведение респондентов. Существуют разные способы ее решения. О них мы уже говорили в п.2.2.2. Это прежде всего группа предложенных западными авторами алгоритмов, в название которых входит аббревиатура AID

(automatic interaction detector). А также некоторые алгоритмы поиска логических закономерностей, предложенные советскими авторами. Отметим, что в этих алгоритмах различны и понятия типа поведения и способы перебора сочетаний значений предикторов.

Наличие сравнительного большого количества алгоритмов, позволяющих решить нашу задачу, объясняется тем, что задача очень актуальна для прикладных исследований (для социологии в частности). За ее решение принимались разные исследователи. И каждый предложил свой подход, свою формализацию соответствующего явления.

Другими словами, мы имеем еще одно подтверждение нашего основного методического положения – для решения практически любой социологической задачи существует несколько методов и, следовательно, на первый план выходит проблема их сравнения, комплексного использования и т.д. Учитывая это, перейдем к рассмотрению конкретных алгоритмов. При этом будем стремиться выделять те их элементы, которые имеют непосредственное отношение к пониманию типа поведения респондентов. Сначала обсудим два известные западные алгоритма.

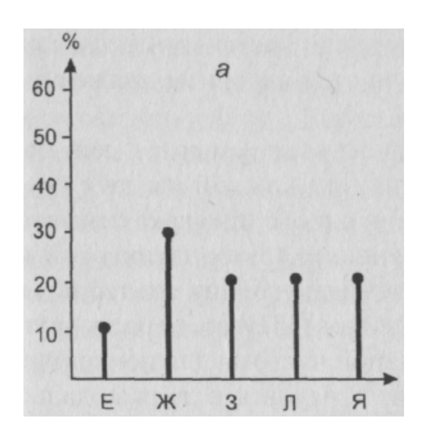
### 2.5.3.2. Алгоритм THAID

*Понимание типа объектов.* Будем считать, что у нас задан некоторый номинальный признак Y – отвечающий, например, рассматриваемому выше вопросу в анкете: За кого Вы собираетесь голосовать? – с 5-ю альтернативами – вариантами ответов: Е, Ж, З, Л, Я. Для каждой проверяемой группы объектов будем

260

вычислять распределение входящих в нее респондентов по этому признаку, подсчитывать соответствующее модальное значение и определять долю его встречаемости. Соответствующий процент будет служить оценкой качества группы с точки зрения возможности рассматривать ее как тип.

Приведем примеры. Предположим, что распределения в каких-то двух группах выглядят следующим образом.



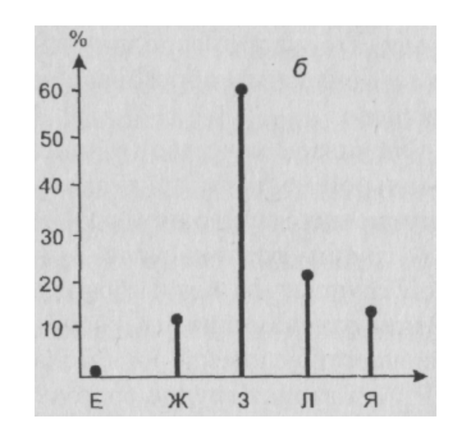




**Рис. 18 Примеры частотных распределений, отражающих электоральное поведение двух групп респондентов**

Модальное значение для первой совокупности – Ж, его доля – 30 %. Для второй же совокупности мода – З. Ее доля – 60%. Качество второй совокупности выше. Однако, вероятно, мы ни ту, ни другую группу не можем рассматривать как тип, поскольку оба процента не достаточно высоки для того, чтобы можно было считать группу "олицетворяющей" определенный тип поведения. Отметим, что содержательные типы тут в принципе будут разными – каждая группа будет ассоциироваться со своим "модальным" политическим лидером.

***Алгоритм перебора сочетаний значений предикторов.*** Как мы уже отметили, алгоритм придуман именно для того, чтобы некоторые сочетания значений предикторов заведомо не просматривались машиной. Социологу важно знать, какие именно. Чтобы это понять, рассмотрим алгоритм.

***Первый шаг.*** Работаем с каждым признаком отдельно. Перебираем следующие варианты разбиения всех его альтернатив на две части: (первая – все остальные); (первая и вторая – все остальные); (первая,вторая, третья – все остальные) и т.д. до последнего варианта: (все, кроме последней, – последняя). Подчеркнем, что перебираются не все возможные варианты сочетаний значений одного признака: множество значения разбивается только на *две* части и "склеиваются" только *соседние* градации. Если мы полагаем, что, например, один тип не могут составлять люди с высшим и начальным образованием, то этот алгоритм должен быть отвергнут.

Оцениваем качество (в описанном выше смысле - как долю модальной частоты признака-функции) каждой из двух групп, получающихся при одном разбиении одного признака (имеются в виду группы респондентов, отметивших альтернативы той или иной группы; мы как бы отождествляем группу альтернатив и группу отвечающих им респондентов). Пусть первая

группа включает $n_1$ человек и доля модальной частоты для нее составляет $P_1$ %, а вторая группа состоит из $n_2$ человек и доля модальной частоты составляет $P_2$ %. Тогда вычислим показатель качества всего разбиения:

$$P = n_1 \times P_1 + n_2 \times P_2$$

Заметим, что здесь мы по существу имеем дело с взвешенным средним. Такой способ усреднения очень распространен в социологии.

Итак, каждое разбиение совокупности альтернатив каждого признака получило свою оценку качества. Выберем наилучшее. Скажем, таковым оказало разбиение совокупности альтернатив признака “образование” на группы (1,2) и (3,4,5). Далее будем изучать респондентов каждой группы отдельно.

***Второй шаг.*** Берем респондентов с низким образованием (отметивших альтернативы 1 и 2, означающие, скажем, начальное и неполное среднее образование) и делаем для них то же самое, что только что делали для всех респондентов (естественно, отличие будет состоять в том, что признак “образование уже не будет

262

рассматриваться). Получим самое хорошее разбиение совокупности респондентов - скажем, это будет разбиение по признаку “семейное положение”, группы альтернатив (1, 2) и (3).

Далее будем изучать отдельно тех людей с низким образованием, которые женаты или неженаты (альтернативы 1 и 2 соответственно) и тех людей с низким образованием, которые разведены (альтернатива 3). И будет это делаться на третьем шагу. А на втором мы должны рассмотреть людей с высоким образованием (отметивших альтернативы 3,4,5 - среднее, неполное высшее и высшее образование соответственно) и реализовать для них ту же процедуру. Допустим, для них наилучшим оказалось разбиение по социальному происхождению, группы альтернатив (1) и (2 и 3). Тогда на третьем шаге мы будем изучать отдельно группы людей с высоким образованием, из семей рабочих (альтернатива 1) и людей с высоким образованием из семей служащих или военных (альтернативы 2 и 3).

Таким образом, у нас уже образовались цепочки, изображенные на рис. 19.

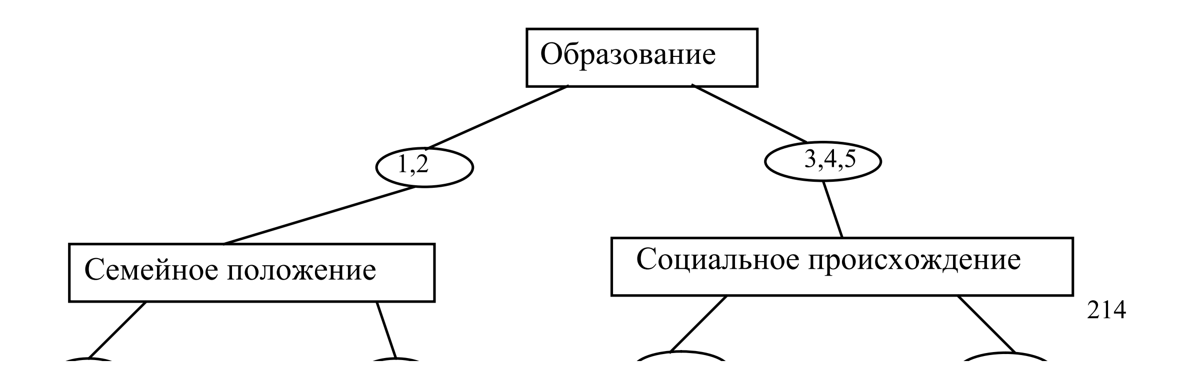


**Рис. 19. Пример результата работы алгоритма THAID**

На ***третьем шаге*** каждая из четырех получившихся групп разделится еще на две. И каждый раз мы будем получать группы с увеличивающейся долей модальной частоты по нашему признаку-функции. Каждую “цепочку” можно считать описанием той группы людей, которая “висит” на конце этой “цепочки”.

263

Чтобы понять,чем дело кончится, перечислим причины останова действия машины. Сразу отметим, что они довольно типичны для анализа социологических данных, действуют при решении очень многих задач, при работе многих, весьма различных алгоритмов.

**Причины останова.**

1) Найдена “хорошая” группа, т.е. такая, в которой упомянутая доля модальной частоты достатоточно велика. Скажем, может оказаться, что среди людей с низким образованием и разведенных 95% проголосовали за Л. Тип найден и крайняя левая нижняя группа в дальнейшей работе не участвует.

2) Получена слишком малочисленная группа. Здесь мы можем поступить по-разному: или игнорировать это обстоятельство и двигаться дальше, исключив соответствующих людей из рассмотрения (как чаще всего и поступают) или попытаться выяснить, в чем состоят те особенности этих людей, изучить их без претензий на статистические обобщения.

3) Получена слишком длинная цепочка. Интерпретация этого обстоятельства очень важна для социолога. Здесь мы имеем дело с пониманием того, что такое та закономерность, которая ищется с помощью любого метода анализа данных. Дело в том, что само понятие закономерности предполагает достаточно простую ее структуру того, что мы закономерностью называем. Слишком длинное описание получающегося типа мы не будем воспринимать как тип. Вряд ли мы сделаем серьезные выводы на основе знания того факта, что люди с высоким образованием, неженатые, живущие в сельской местности, имеющие более 4-х детей, 3-х поросят, не любящие смотреть телевизор и мечтающие о путешествии на Кипр почти все проголосовали за Л. Причинно-следственные закономерности останутся за бортом наших рассуждений. (По той же причине мы обычно не воспринимаем как закономерность

классификацию, в которой 1500 классов или результат факторного анализа, которых дал нам 150 латентных переменных.) Об этом мы говорили в п.1.4 части I

4) ЭВМ не нашла ни одной совокупности с интересующими нас свойствами. В рассматриваемом примере - ни одной группы

264

респондентов, среди членов которой интересующего нас мнения придерживалась бы достаточно большая доля людей. Это означает то, что в используемой анкете не заложено описание интересующего нас поведения. Такая ситуация может быть следствием нашего неумения составлять анкету, общаться с респондентом, учитывать цели исследования при формировании инструментария, ставить задачу и т.д.

Подводя определенный итог, можно сказать, что задача поиска детерминирующих сочетаний значений предикторов может пониматься как единство трех задач: (1) выделение из числа независимых переменных наиболее информативных в том смысле, что именно по сочетанию их значений с наибольшей степенью уверенности можно судить о типе поведения объектов; (2) выяснение, какие именно сочетания значений информативных признаков детерминируют указанный тип (в том числе то, какие из этих значений должны объединяться "склеиваться"); (3) выявление конкретных типов поведения, свойственных объектам рассматриваемой совокупности (т.е. конкретных характеризующих выделяемые группы модальных значений, встречающихся с достаточной частотой; ясно, что, скажем, далеко не для каждого кандидата, вообще говоря, найдется "его" группа респондентов).

Рассмотренный алгоритм задействован в известном западном пакете OSIRIS. Коротко описание этого подхода можно найти в [Интерпретация и анализ ..., 1987. С.29, с.136-151; Рабочая книга ..., 1983. С. 193-195; Типология и классификация ..., 1982. С.213-230]. Там он называется также алгоритмом последовательных разбиений. См. также литературу, указанную в п. 2.2.2. Отметим также, что буквы TH в начале имени алгоритма означают греческую букву $\vartheta$, поскольку именно так обозначили авторы алгоритма тот связанный с долей модальной частоты критерий качества выделяемых групп респондентов, который мы описали выше.

### 2.5.3.3. Алгоритм CHAID

Как и при работе алгоритма THAID, задается номинальный признак-функция Y. Поведение каждого респондента здесь понимается так же, как выше (скажем, это выбор респондентом той или иной позиции при голосовании). А вот групповое поведение будем

оценивать по-другому. А именно, будем ассоциировать его не с частотой модального значения признака Y, а со всем распределением этого признака. Как и выше, в нашу задачу, наряду с поиском сочетаний значений рассматриваемых признаков, детерминирующих интересующее нас групповое поведение, входит поиск конкретных видов такого поведения - конкретных распределений значений признака Y, детерминируемых нашей анкетой.

Алгоритм состоит из ряда шагов, сходных с теми, которые были описаны выше. На каждом шаге происходит склеивание определенных градаций каждого признака и выделение той переменной, в соответствии со значениями которой совокупность респондентов делится далее на части.

Рассмотрим принципиальные моменты алгоритма, связанные с пониманием искомых типов поведения респондентов и позволяющие реализовывать упомянутые процедуры.

**Определение склеиваемых градаций**. Покажем на примере, как определяется, какие градации анализируемого признака X должны склеиваться.

Пусть Y – электоральное поведение респондента в том же смысле, какой был использован в п. 2.5.3.2, а признак X – это профессия с градациями “врач”, “учитель”, “рабочий”. Рассмотрим частотную таблицу, связывающую эти два признака (таблица 27).

*Таблица 27.*

**Таблица сопряженности, использованная для определения “склеиваемых” градаций признака “профессия” в процессе использования алгоритма CHAID**

| Профессия | Предполагаемое голосование | | | | | Итого |
|---|---|---|---|---|---|---|
| | Е | Ж | З | Л | Я | |
| Врач | 10 | 2 | 10 | 8 | 30 | 60 |
| Учитель | 5 | 1 | 5 | 4 | 15 | 30 |
| Рабочий | 0 | 30 | 8 | 20 | 2 | 60 |
| Итого | 15 | 33 | 23 | 32 | 47 | 150 |

Склеить мы должны такие градации, которые не имеет смысла рассматривать дальше отдельно из-за того, что респонденты, отметившие одну градацию, обладают тем же электоральным “поведением”, что и респонденты, отметившие другую. Рассмотрение соответствующих совокупностей респондентов отдельно не имеет смысла. Нетрудно видеть, что такими свойствами обладают градации “врач” и “учитель”. Если мы рассмотрим отдельно представителей этих профессий, то уж никак не получим разные типы избирателей: половина

врачей хочет голосовать за Я и половина учителей - тоже. Одинаковое количество учителей (5 человек, примерно 17 %) хочет голосовать за Е и З соответственно, и то же самое можно сказать о врачах и т.д. Нетрудно видеть, что сказанное является следствием того, что первые две строки нашей частотной таблицы пропорциональны.

Относительно же врачей и рабочих мы подобные выводы сделать не можем. Вероятно, эти альтернативы нельзя объединять. Напротив, имеет смысл разделить нашу совокупность на две части, рассмотрев врачей и рабочих отдельно. Они являют собой совершенно разный тип электорального поведения: за Я собираются голосовать 50% (30 человек) врачей и менее 2% (2 человека) рабочих и т.д. Ясно, что это – следствие сильного отклонения от пропорциональности первой и третьей строк нашей таблицы.

Вспомним теперь критерий “хи-квадрат”. Пропорциональность строк таблицы сопряженности означает равенство этого критерия нулю и, следовательно, влечет за собой принятие нуль-гипотезы – гипотезы об отсутствии связи между переменными. Отсутствие пропорциональности влечет отвержение нуль-гипотезы, т.е. согласие с наличием связи между переменными. И приведенные выше рассуждения по существу говорят о том, что склеивать надо те альтернативы, которые, будучи “вырванными” из общего списка и рассмотренные отдельно, как значения “вспомогательного” дихотомического признака (в нашем случае - признака с двумя альтернативами: “учитель” и “врач”) приведут нас к выводу об отсутствии связи между этим вспомогательным признаком и Y.

Но эта формулировка не очень корректна, поскольку критерий “хи-квадрат” не “говорит” о том, есть или нет связь между переменными, а лишь дает основание принять или отвергнуть гипотезу об отсутствии связи на определенном уровне значимости $\alpha$. Поэтому более грамотной будет следующее правило, по которому мы определяем, какие именно две альтернативы рассматриваемого признака надо склеить.

Для конкретного признака X проверяем все пары альтернатив. Считаем, что каждая пара отвечает своему дихотомическому признаку и, задавшись уровнем значимости (скажем, $\alpha$ = 0,05), вычисляем критерий “хи-квадрат” для этого признака и Y. Отбираем те пары, для которых значение $X^2$ не превышает соответствующее критическое значение. Ясно, что это пары, для которых имеет смысл принять нашу нуль-гипотезу. Далее выбираем ту пару, для которой $X^2$ меньше всего, т.е. для которой наша нуль гипотеза принимается как бы с большей надежностью. Именно альтернативы этой пары мы и склеиваем.

**Выбор признака для разбиения совокупности**. Склеив какие-то альтернативы в каждом из анализируемых признаков, мы вычисляем критерий “хи-квадрат” между каждым из оставшихся к рассматриваемому шагу признаком $X_i$ и Y. Здесь поступим противоположным образом по сравнению с тем, что было выше: отберем те признаки $X_i$, для которых наш критерий превышает критическое значение, т.е., для которых имеет смысл отвергнуть гипотезу об их независимости от Y, т.е. считать, что между каждым из них и Y есть связь. Среди этих признаков отберем тот, для которого $\chi^2$ имеет наибольшее значение, т.е. тот, для которого связь существует с наибольшей вероятностью. По его градациям мы и будем далее разбивать совокупность респондентов.

Описанные процедуры мы реализуем так же по шагам, как и в алгоритме THAID. В итоге выделяются группы респондентов, каждая из которых описывается последовательностью значений рассматриваемых признаков (так, последовательность, отвечающая крайней правой “цепочке” с рисунка 19, состоит из двух элементов: среднее, неполное высшее или высшее образование; из служащих или военных). Наш алгоритм дает основание полагать, что каждой из таких выделенных последовательностей будет отвечать свое “поведение” соответствующей группы респондентов, т.е. свое, характерное именно для данной группы, распределение признака Y.

Заметим, что алгоритм CHAID, так же, как и THAID, не гарантирует выявления в исходных данных всех интересующих исследователя закономерностей. Основная причина – в том, что на каждом шаге разбиения алгоритм оценивает лишь двумерную связь. Он может заставить исследователя исключить из дальнейшего рассмотрения такой признак-предиктор, который, будучи сам по себе не очень “хорошим”, в сочетании с другими может дать наилучший результат. Скажем, некий предиктор, не имея связи с целевым и, в силу этого, отбрасываемый (из-за того, что условные распределения целевого признака, вычисленные для отдельных градаций предиктора, схожи друг с другом и поэтому не дают нам отдельные типы респондентов), в сочетании с каким-то другим предиктором может иметь сильную связь с целевым (в п. 2.3.6 мы приводили пример, когда связь между двумя не связанными признаками появляется при фиксации значения третьего признака). И эта связь может быть более значимой, чем связь между целевым признаком и отобранными алгоритмом предикторами.

Алгоритм задействован в известном пакете программ SPSS. Буквы “CH” в названии алгоритма – от греческой буквы “Х” (Хи), поскольку критерий “Хи-квадрат” лежит в основе метода.

Отметим, что описанные алгоритмы охватывают не все те задачи поиска взаимодействий, которые интересуют социолога. Имеются другие направления анализа данных, включающие в себя несколько иные алгоритмы интересующего нас плана - алгоритмы поиска логических закономерностей, разработанные советскими авторами. Об этих алгоритмах пойдет речь в п.п. 2.5.5 и 2.5.6.

### 2.5.4. Методы ДА, THAID, CHAID с точки зрения поиска обобщенных взаимодействий

Вспомним расширенное, обобщенное определение понятия взаимодействия из п.2.2.1 и рассмотрим, в какой мере рассмотренные алгоритмы позволяют находить такие обобщенные взаимодействия. Вспомним также те примеры выводов в терминах изучаемых признаков, которые мы привели в названном параграфе, считая, что именно они в основном интересуют социолога.

Начнем с рассмотрения ДА. Ясно, что он направлен на поиск таких сочетаний значений предикторов, которые действительно можно назвать взаимодействиями. Он позволяет получать истинные суждения такого типа: “5-е или 6-е значение 8-го признака в сочетании с 3-м значением 14-го и 1-м значением 2-го детерминирует 2-е значение 30-го". Однако очевидно, что при этом имеются в виду не все наши обобщенные взаимодействия. Не учитываются следующие обстоятельства.

(1) В обобщенном определении взаимодействия в качестве объясняющего положения может выступать *любая* логическая функция от значений исходных признаков. Помимо конъюнкции и дизъюнкции, задействованных в ДА, могут использоваться отрицание и импликация. Это в какой-то мере не принципиально, поскольку функции второй пары в нашем случае могут быть выражены через функции первой, но социологу при формулировке содержательных задач часто бывает легче, естественнее использовать все элементарные функции логики высказываний. Например, предположим, что вопрос о занятии респондента предусматривает 15 ответов: токарь, пекарь, …, аптекарь, бомж. Наверное, исследователю удобнее проверять истинность суждения "если респондент – не бомж, то он согласен на оплату благоустройства дворов", чем суждение "если респондент или токарь, или пекарь, или …, или аптекарь, то он согласен на оплату …");

(2) При использовании ДА в качестве объясняемого положения выступает некоторое единственное значение какого-либо независимого признака При расширенном же определении

взаимодействия, в соответствии с нашим определением, объясняемым положением может служить также любая логическая функция от сочетаний значений одного или нескольких признаков, некоторым другим образом задаваемое "поведение" респондента (см. ниже обсуждение алгоритма CHAID), частота таблицы сопряженности; кроме того, предусматривается возможность отсутствия объясняемого положения. Всего этого ДА не учитывает.

270

Перейдем к рассмотрению алгоритмов THAID и CHAID. Нетрудно видеть, что они, как и ДА, направлены на поиск взаимодействий. Но здесь тоже учитываются не все свойства наших обобщенных взаимодействий. Названные алгоритмы позволяют делать выводы такого плана:

“5-е или 6-е значение 8-го признака в сочетании с 3-м значением 14-го и 1-м значением 2-го детерминирует групповое поведение, описанное (в определенном в п. 2.5.3 смысле) в терминах 30-го признака”. Для алгоритма THAID упомянутое “поведение” означает долю модального значения 30-го признака. Выделенные группы – те, для которых эта доля достаточно высока. Для алгоритма CHAID – “поведение” характеризуется распределением выходного (в данном случае – 30-го) признака. Выделенные группы таковы, что отвечающие им распределения максимально отличаются друг от друга.

По поводу объясняющего положения, фигурирующего в обоих алгоритмах, можно сказать то же, что было сказано выше применительно к возможностям ДА.

Переходя к обсуждению объясняемого положения, рассмотрим сначала алгоритм THAID. Цели ДА здесь достигаются. Это является следствием того, что обеспечение максимальной (из возможных) доли модального значения выходного признака по существу означает обеспечение того, что соответствующее объясняющее выражение детерминирует это самое модальное значение. Преимуществом алгоритма THAID является определенная гарантия того, что, если искомые детерминации существуют в исследуемой совокупности, то они будут выявлены. Кроме того, THAID позволяет не “замыкаться” на единственном значении выходного признака, а искать все такие его значения, для которых можно найти соответствующее объясняющее выражение.

Пока мы говорили о возможности пропустить интересующие исследователя факты. Теперь попытаемся сравнить сами критерии качества детерминаций. Другими словами, сравним способы формализации понятия приближенности связи между объясняющим и объясняемым положениями в рассматриваемых ситуациях.

В ДА упомянутый способ формализации – это точность и полнота строящихся детерминаций. В случае использования THAID степень приближенности найденных детерминаций определяется выбором пороговой доли модальной частоты целевого признака. Такая доля – это “точность правила” в смысле ДА. А поскольку мы при использовании THAID ищем сразу все достаточно точные детерминации, то можно сказать, что в результате нами находятся и достаточно полные правила. Объясняющие положения, отвечающие одному и тому же объясняемому значению выходного признака, при этом объединяются в дизъюнкцию.

Таким образом, в принципе THAID позволяет решать те же задачи, что и ДА, но с большей эффективностью. Явным преимуществом ДА является то, что здесь мы активно используем интуицию исследователя. Это обстоятельство может существенно восполнить сформулированные в п.2.5.3.2 недостатки алгоритма THAID, приводящие к определенным “проколам” в его работе, к пропуску части искомых сочетаний значений предикторов.

При использовании алгоритма CHAID объясняемое положение – это такое "поведение" объектов выделенной группы, которое отождествляется с характерным только для нее распределением целевого признака. Подчеркнем, что такое "поведение" в принципе отличается от того, что было обсуждено выше. При использовании ДА и THAID поведение определяется одним значением выходного признака. Это значение выступает как вполне самостоятельная сущность, описывающая что-то важное для социолога. При использовании же CHAID выходной признак предстает перед нами целиком, в виде вероятностного (частотного) распределения. Здесь мы явно имеем дело с той группой методов, которая в п. 2.2.3 связывалась нами с существованием числовых латентных переменных, стоящих за наблюдаемыми номинальными признаками. Это предполагает само использования критерия “Хи-квадрат”.

Ясно, что и при использовании CHAID учитываются не все требования, фигурирующие в нашем обобщенном определении взаимодействия. Не учитывается, что в качестве объясняемого положения может быть логическая функция от значений одного или нескольких признаков, частота таблицы сопряженности и то, что объясняющее положение может отсутствовать. Последнее обстоятельство будет рассмотрено в следующих двух параграфах. Там речь пойдет о проверке истинности некоторой логической формулы.

Алгоритм CHAID тоже не гарантирует получения всех интересующих исследователя решений. Более того, он не всегда позволяет повышать качество выделяемых типов объектов. Об этом шла речь в п. 2.5.3.3. Тем не менее, он как и THAID, все же в большей мере позволяет осуществлять целенаправленный поиск закономерностей, чем это делает ДА.

### 2.5.5. Поиск логических закономерностей: элементы исчисления высказываний; понятие закономерности; алгоритм поиска; его сравнение с ДА.

Направление, о котором пойдет речь, отражает достижения новосибирских ученых. Оно включает в себя очень много разработок, начиная с полуфилософских размышлений о том, что такое закономерность, и кончая огромным количеством алгоритмов, позволяющих искать конкретные закономерности различной степени общности [Витяев Е.Е., Логвиненко А.Д., 1999; Загоруйко, 1979; Лбов, 1981; Рабочая книга ..., 1983. С.197-198]. . Мы полагаем, что эти разработки достойны внимания социологов. Приходится сожалеть, что российские исследователи, активно пользуясь западными пакетами и, следовательно, западной методологией анализа данных, зачастую не знают работ соотечественников. А их достижения при решении многих задач в большей степени отвечают естественной логике социолога и во многом более надежны.

Мы лишь очень коротко коснемся соответствующих проблем. Следуя авторам цитируемых работ, введем понятие логических закономерностей (и тем самым еще раз покажем, что решение широкого круга социологических задач требует использования специфического языка – языка математической логики). При этом рассмотрим лишь один их вид и один из простейших алгоритмов их поиска.

#### *Элементы исчисления высказываний.*

Прежде, чем строго определить понятие *логической закономерности,* необходимо ввести несколько вспомогательных определений. Это даст нам возможность не только описать один из конкретных алгоритмов поиска логических закономерностей, но и более строго говорить о том, о чем шла речь в предыдущих параграфах.

Пусть $X_1$, $X_2$, ..., $X_m$, $Y_1$, $Y_2$, ..., $Y_n$ – какие-то изучаемые нами признаки. Назовем *элементарными высказываниями (суждениями)* выражения вида: ($X_2 = 5$); ($3 \le X_n \le 5$) (такого рода высказывания здесь нас не интересует, поскольку они касаются порядковых шкал, а мы рассматриваем только номинальные признаки, но порядковые шкалы, вообще говоря, конечно, отнюдь не безынтересны для социолога; поэтому мы не будем сокращать изложение цитируемых авторов за счет ликвидации всего, что с ними связано); ($Y_4 = 34{,}2$) и т.д.

Будем продолжать считать, что читателю знакомы логические связки $\rceil$, &, $\vee$, $\supset$ (отрицание, конъюнкция, дизъюнкция, импликация) и отвечающие им таблицы истинности, и

введем определение логической *формулы*, являющееся ключевым для математической логики и принадлежащее тому ее разделу, который носит название "исчисление высказываний". Определение рекурсивно:

1) все элементарные суждения суть формулы;

2) если $F_1$ и $F_2$ – формулы, то и $(\neg F_1)$, $(F_1 \& F_2)$, $(F_1 \vee F_2)$, $(F_1 \supset F_2)$ – формулы;

3) других формул, кроме тех, что получаются в соответствии с предыдущими пунктами, не существует.

Ниже формулы будем называть также *суждениями* или *высказываниями*.

Теперь приведем рекурсивное определение *длины формулы*:

1) Все элементарные суждения и их отрицания имеют длину, равную единице;

2) Если формула $F_1$ имеет длину m, а формула $F_2$ – длину n, то формулы $(F_1 \& F_2)$, $(F_1 \vee F_2)$, $(F_1 \supset F_2)$ имеют длину $(m + n)$.

Описание языка математической логики (в рамках т.н. узкого исчисления предикатов) будет продолжено в п. 2.5.6).

***Логические закономерности, характеризующие заданный класс объектов.***

Рассмотрим задачу, имеющую более широкий характер, чем те, что были рассмотрены в предыдущих параграфах – задачу описания какого-либо класса объектов.

Предположим, что нас интересует, как в терминах наших признаков нужно описать некоторый класс объектов ω. В качестве такого класса может служить любое множество респондентов с изучаемым социологом "поведением" объектов, Например, это может быть класс респондентов, проголосовавших за политического лидера А. Именно этот класс ниже будет рассматриваться в качестве примера.

По существу, говоря о классе ω, мы имеем в виду какое-либо из тех множеств, которые выше у нас ассоциировались с выполнением объясняемого выражения. Если объясняемое выражение – формула, то можно сказать, упомянутый класс – это совокупность объектов (респондентов), на которых эта формула выполняется. И задача описания класса – это задача поиска объясняющего выражения. Правда, здесь имеется отличие от тех представлений об объясняющем выражении, которые использовались выше. Предположим, например, что мы выяснили, что все женщины – жители села старше 70 лет голосуют за рассматриваемого кандидата. Тогда соответствующее сочетание значений признаков можно считать объясняющим выражением в смысле ДА, а также алгоритмов THAID или CHAID. Но оно вполне может быть

отвергнуто как выражение, описывающее рассматриваемый класс, если окажется, что доля старых жительниц села среди всех проголосовавших за нашего кандидата очень мала. Другими словами, от искомых выше детерминирующих выражений мы в первую очередь требовали точность (интенсивность), а здесь мы даже очень точное выражение отвергнем, если у него малая полнота (емкость). Однако, как мы увидим ниже, соответствующую планку (полноту) в рассматриваемых в настоящем параграфе алгоритмах можно делать как угодно низкой. Поэтому в принципе с их помощью могут быть выявлены любые точные взаимодействия. К обсуждению этого вопроса мы вернемся в конце параграфа, а сейчас приступим к описанию одного из алгоритмов, позволяющих найти описание априори заданного класс объектов.

Будем говорить, что *логическая формула выполняется на некотором объекте* (объектами у нас чаще всего являются респонденты и в таком случае говорят о выполнении формулы для респондента), если эта формула истинна для этого объекта. Выше мы фактически использовали это определение, не вводя его строго, формально. К примеру, используя фразу: “ 5-е значение 8-го признака часто встречается с 3-м значением 14-го и 1-м значением 2-го”, мы имели в виду то, что выполнение формулы ($X_8 = 5$) для некоторого респондента часто сочетается с выполнением для него же формулы (($X_{14} = 3$) & ($X_2 = 1$)).

Обозначим через $\varpi$ совокупность объектов, не принадлежащих этому классу.

Зададимся некоторыми критериями $\alpha$ и $\beta$, изменяющимися от 0 до 1, но разными по величине: $\alpha$ – достаточно большое (скажем, больше 0,8), а $\beta$ – достаточно малое (скажем, меньше 0,1).

Назовем некоторое суждение **s** *логической закономерностью, характеризующей класс* $\omega$, если это суждение выполняется для достаточно большой доли элементов этого класса и для достаточно малой доли элементов $\varpi$. При этом достаточно большой долей будем называть такую долю p, для которой выполняется неравенство

$p \geq \alpha$,

а достаточно малой – такую долю q, для которой справедливо соотношение

$q \leq \beta$.

Ясно, что любая закономерность, характеризующая класс $\omega$, может служить его описанием. Покажем, как можно искать такие описания. Кратко опишем один из самых простых алгоритмов – алгоритм ТЭМП [Лбов, 1981. С.40-41].

Будем считать, что у нас заданы описанные выше критерии $\alpha$ и $\beta$, т.е. определено, какую логическую формулу можно называть закономерностью, а какую – нельзя. Для произвольного

высказывания s обозначим через $p_{s\omega}$ долю тех объектов из ω, для которых выполняется s, а через $p_{s\varpi}$ аналогичную долю объектов из $\varpi$.

В качестве примера при описании алгоритма рассмотрим ситуацию, когда требуется выявить "портрет" респондента, голосующего за кандидата А. Пусть в анкете имеется три вопроса: $X_1$ – пол (1 – мужчина, 2 – женщина), $X_2$ - место жительства (1 – крупный город, 2 – небольшой город, 3 – село), $X_3$ – образование (1 – начальное, 2 – неполное среднее, 3 – среднее, 4 – высшее). Таким образом, в нашем примере ω – это класс голосующих за кандидата А, описание класса – это "портрет" составляющих его респондентов в терминах указанных признаков.

Алгоритм ТЭМП представляет собой некий перебор высказываний. При этом в качестве тех логических функций, в виде которых ищется искомая закономерность, используются только конъюнкции. Это существенно для понимания алгоритма. Опишем этапы предусматриваемого алгоритмом перебора суждений.

1. Рассмотрим все элементарные высказывания и их отрицания, т.е. все формулы длины 1. Для каждого высказывания **s** проверяем выполнение условия $p_{s\omega} \geq \alpha$. Если условие не выполняется, то высказывание исключается из дальнейшего рассмотрения. Если выполняется, то проверяем выполнение условия $p_{s\varpi} \leq \beta$. Если и это условие выполняется, то считаем, что **s** – одна из искомых закономерностей и выдаем ее на печать. Если $p_{s\varpi} > \beta$, то высказывание **s** запоминается и сохраняется в памяти машины. Такие высказывания далее будем называть *отмеченными.*

Предположим, к примеру, что, рассмотрев формулы вида ($X_1 = 1$), ($X_1 = 2$), …, ($X_2 = 1$), ($X_2 = 2$), …, мы выяснили, что среди проголосовавших только доли лиц, обладающих свойствами ($X_1 = 1$) (т.е. доля мужчин), ($X_2 = 3$) (т.е. доля жителей села), ($X_3 = 3$) и ($X_3 = 4$) (доли лиц с высшим и средним образованием) больше α. Именно эти свойства и служат основанием для дальнейшего поиска закономерностей. Остальные свойства отбрасываем. Ведь если, скажем, доля женщин в рассматриваемом классе меньше установленного нами порога, то таковой будет и доля женщин, проживающих в селе, и доля женщин с начальным образованием и т.д. Другими словами любая конъюнкция, одним из элементов которой будет служить выражение ($X_1 = 2$), заведомо будет выполняться для очень малого количества объектов нашего класса и, вследствие этого? заведомо не будет закономерностью. Значит, женщины в принципе должны быть исключены из дальнейшего рассмотрения.

Далее проверим, какое количество респондентов, *не* проголосовавших за А, обладает отобранными свойствами. Предположим, что доля мужчин, т.е. людей со свойством ($X_1 = 1$) оказалась здесь меньше нашего порога $\beta$. Это значит, что указанное свойство – одна из искомых закономерностей: доля мужчин среди проголосовавших за А достаточно велика, а среди непроголосовавших – достаточно мала. Мы это учитываем и далее свойство "быть мужчиной" исключаем из рассмотрения. Это разумно, поскольку в данной ситуации вряд ли нам даст что-то новое отдельное изучение, скажем, мужчин – селян или мужчин со средним образованием. Некоторые из свойств подобного рода вполне могут удовлетворять нашему определению закономерности. Исключая свойство "быть мужчиной" из дальнейшего рассмотрения, мы тем самым обеспечиваем получение закономерностей минимальной длины: ни одно высказывание, получаемое из закономерности путем исключения любого элементарного высказывания, не будет уже закономерностью.

Пусть теперь оказалось также, что доля жителей села среди непроголосовавших больше $\beta$. Значит, свойство "быть жителем села", т.е. ($X_2 = 3$), не является закономерностью. Но оно может стать таковой в сочетании с какими-то другими свойствами. Значит, мы это свойство должны оставить для дальнейшей работы, сделать его отмеченным. Пусть также отмеченными будут и свойства ($X_3 = 3$) и ($X_3 = 4$)

2. Второй этап работы состоит в рассмотрении конъюнкций всех суждений, отмеченных на первом этапе. Рассуждения аналогичны описанными выше: если для какого-то суждения **s** не выполняется условие $p_{s\omega} \geq \alpha$ (т.е. если $p_{s\omega} < \alpha$), то суждение исключается из дальнейшего рассмотрения. Последнее справедливо и для составляющих его элементарных высказываний. Если условие $p_{s\omega} \geq \alpha$ выполняется, то проверяем справедливость условия $p_{s\bar{\omega}} \leq \beta$. При его справедливости суждение считается одной из найденных закономерностей и выдается на печать. При невыполнении условия $p_{s\bar{\omega}} \leq \beta$ (т.е. при $p_{s\bar{\omega}} > \beta$) оба составляющих s элементарных высказывания отмечаются и оставляются в памяти.

3. Рассматриваются всевозможные конъюнкции длины три с аналогичной проверкой указанных условий и т.д.

Нетрудно видеть, что описанный алгоритм позволяет обнаружить все закономерности, "скрывающиеся" в исходных данных. Более того, как мы уже упоминали, рассматривая первый шаг, найденные закономерности представляют собой высказывания минимальной длины – ни из одной закономерности нельзя выкинуть никакой составляющей ее подформулы без того, чтобы закономерность не перестала быть закономерностью.

*Сравнение рассмотренного алгоритма с ДА.*

Как мы уже упоминали в начале параграфа, алгоритм ТЭМП ориентирован на поиск как можно более полных детерминаций. Однако при умелом регулировании величин $\alpha$ и $\beta$ можно гарантировать и нахождение всех точных детерминаций. Если мало $\alpha$, мы будем получать закономерности, справедливые и для малых долей элементов $\omega$. Если же достаточно малым будет и $\beta$, мы можем достичь того, чтобы среди объектов, принадлежащих $\varpi$, практически не было таких, на которых наша закономерность выполняется. Это означает, что соответствующая детерминация будет точной относительно $\omega$.

Таким образом, алгоритм ТЭМП не только позволяет решать задачи, решаемые с помощью ДА, но и дает возможность делать это более эффективно, с гарантией того, что нами были выявлены *все* интересующие нас закономерности (взаимодействия). Более того, рассмотренный в настоящем параграфе подход дает возможность широко варьировать то, **что** детерминирует наше взаимодействие: множество $\omega$ произвольно. Кроме того, некоторое преимущество описанного алгоритма заключается в возможности использования таких формулировок искомых закономерностей, которые включают в себя отрицания элементарных суждений. Мы уже отмечали, что это зачастую бывает удобно для социолога.

Алгоритм ТЭМП – это лишь один из самых простых алгоритмов, лежащих в русле того мощного подхода к поиску эмпирических закономерностей, который был предложен новосибирскими учеными [Загоруйко, 1979; Лбов, 1981]. В рамках этого подхода может решаться гораздо более широкий круг важных для социолога задач, чем тот, которого мы касаемся. Этот круг включает в себя, помимо задач поиска логических закономерностей задачи распознавания образов, поиска эффективной системы признаков, эмпирического предсказания и т.д.

Отметим, что пакеты программ, реализующие предложенные новосибирскими учеными методы поиска логических закономерностей, разработаны в ИМ СО РАН (например, пакет ОТЭКС).

### 2.5.6. Поиск логических закономерностей и теория измерений. Элементы узкого исчисления предикатов

В настоящем параграфе пойдет речь о разработках, позволяющих связать проблему нахождения взаимодействий с проблемой измерения. Подчеркнем, что этот параграф отличается от других тем, что здесь мы не будем рассматривать конкретные примеры реализации затрагиваемых положений (это сложно, требует приобщения читателя к достаточно серьезным утверждениям математической логики), а затронем эти положения на теоретическом уровне, указав тем самым направление, представляющееся нам перспективным для социологической практики.

Сначала – некоторые предварительные замечания о естественности и актуальности постановки вопроса о связи репрезентационной теории измерений (РТИ) и анализа данных .

Вспомним, что анализу данных предшествует этап измерения, представляющий собой выбор (построение) эмпирической и математической систем (ЭС и МС) и адекватное отображение первой во вторую (п.2.2 части I). Все наши методы – это некие способы изучения той МС, которая, как мы считаем, является хорошей моделью выявленной на первой стадии исследования ЭС. Найденные закономерности интересуют нас прежде всего как свойства ЭС (здесь мы не анализируем проблемы получения самой ЭС, что тоже – весьма непростое и творческое дело; см. первую часть книги). И представляется естественным стремление выявить, не могут ли некоторые утверждения РТИ способствовать более эффективному изучению ЭС. С другой стороны, некоторые соображения позволяют надеяться, что соответствующие изыскания могут быть весьма полезными, поскольку при "измеренческой" постановке вопроса о поиске статистических закономерностей мы имеем возможность проанализировать главные причины, вызывающие сложности использования в социологии математического аппарата, - причины, связанные с успешностью моделирования, отражения реальности в математических конструктах (ведь измерение – это и есть моделирование такого рода). Об этих сложностях мы неоднократно говорили в первой части работы.

Итак, попытаемся привлечь достижения РТИ для получения более адекватных выводов о структуре изучаемой ЭС.

В настоящей работе рассматривается лишь один вид МС – т.н. признаковое пространство (кстати, – это нечисловая система). Для нас осуществление измерения – это окончательный переход к мышлению признаками.

Далее, анализируя полученные в результате измерения данные, мы будем получать “содержательные” выводы, формулируемые в терминах признаков. О том, какой вид эти выводы могут иметь, мы много говорили в предыдущих параграфах. Надеемся, что освоившему их читателю ясно, что главное, что требуется решить для обеспечения соответствия выводов

реальности и достаточной широты совокупности этих выводов (последнее – для того, чтобы можно было говорить о хорошем изучении ЭС) – это проблема выбора адекватного метода. Выше, сравнивая разные алгоритмы, мы показывали, что результаты, полученные разными методами, могут отличаться друг от друга, что закономерности, четко

выделяемые одним методом, могут быть "не замечены" другим и т.д. Существуют ли вполне адекватные реальности способы выявления закономерностей?

Проблема адекватности метода относительно легко решается при соблюдении классического для РТИ соотношения вида

ЭСО — homo → ЧСО (6)

Во всяком случае, тут ясно, в каком направлении можно обосновывать эту адекватность (мы имеем в виду устойчивость результата применения метода относительно допустимых преобразований используемых для получения исходных данных шкал; об этом см. литературу по РТИ [Суппес и Зинес, 1967; Толстова, 1998; Krantz et all., 1971-1990] ).

Однако для того, чтобы говорить о выборе адекватного метода поиска статистических закономерностей в “измерительном” ракурсе, необходимо вспомнить, что потребности практики уже давно потребовали от теории измерений определённых обобщений тех классических представлений, о которых шла речь выше. В [Толстова, 1998] отмечалось, что жизнь заставляет социолога отказаться и от принятия в расчёт только тех эмпирических отношений, которые значимы для традиционно рассматриваемых типов шкал, и вообще от задания эмпирической системы (ЭС) в виде системы с отношениями, и от понимания шкалы как гомоморфизма, и от трактовки измерения как отображения реальности именно в числовую систему, а не в произвольную математическую (с отношениями или без - соответственно, МСО и МС). Развиваясь, РТИ заставляет нас выделить главное в ней - понимание измерения как построения модели ЭС с помощью элементов некой МС – и именно это положить в основу всех рассмотрений.

Заметим, что, если рассматривать проблему адекватности метода как проблему выбора модели изучаемого явления (именно о такой модели идет речь, когда мы, например, используем тот или иной коэффициент связи, или определяем, скажем, каким алгоритмом – THAID или CHAID – пользоваться при склеивании градаций какого-либо признака), то имеет смысл свойства эмпирических объектов, вытекающие из справедливости выбираемой модели, считать свойствами ЭС. Наличие таких свойств и является основной причиной, заставляющей нас отступать от соотношения (6). Подчеркнем, что подобные рассуждения приводят нас к

необходимости рассматривать все социологическое исследование как некий обобщенный процесс измерения.

Ответ на вопрос о том, что делать, если схема (6) неверна, если жизнь выводит исследователя за ее рамки, зависит от того, каковы причины нарушения схемы. Рассмотрим одну из ситуаций, когда ЭС не удается задать в виде системы с отношениями, но удается как-то ее формализовать с помощью введения определенной аксиоматики.

Некоторые аспекты, связанные с возможностью аксиоматического определения ЭС настолько важны, что им было уделено большое внимание многими исследователями. Объём и практическая значимость соответствующих разработок позволяют говорить о рождении специфического аксиоматико-репрезентационного подхода к пониманию измерения (Axiomatic-Representational Viewpoint in measurement) [Krantz et al., 1990. P.201]. Ниже мы на примере продемонстрируем тот аспект, разработка которого принадлежит российским ученым [Витяев, Логвиненко, 1998]. Сформулируем некоторые принципы, изложенные в названной работе, попытавшись параллельно показать, как они вписываются в некоторую более широкую картину современного положения дел с анализом социологических данных.

Рассматриваемый подход вносит существенные дополнения в РТИ: он направлен не на привнесение аксиом в ЭС из каких-либо внешних по отношению к ней соображений (конечно, с последующей их проверкой на ЭС, проверкой, использующей принцип фальсифицируемости и теорию статистического вывода), а на выявление ("открытия") этих аксиом из анализа самой ЭС. В основе подхода лежит представляющаяся весьма полезной постановка вопроса о том, нельзя ли каким-либо конструктивным способом описать всё множество содержательных выводов, которые могут быть получены для конкретной совокупности полученных социологом данных. Оказывается, что этот вопрос не бессмыслен, на него существует положительный ответ. Именно об этом и пойдёт речь. Прежде всего расширим тот логический язык, который выше был использован для описания основных интересующих социолога закономерностей. А именно, от языка исчисления высказываний (описанного в п.2.5.5) перейдем к языку исчисления предикатов первого порядка.

### *Описание языка узкого исчисление предикатов*

Опишем соответствующий алфавит для рассматриваемого случая. Прежде всего – о нелогических символах формализованного языка.

Предметные (индивидные) константы: конкретные номера респондентов, для обозначения которых могут использоваться буквы a, b, c,... . Предметные (индивидные) переменные – обозначения произвольных номеров респондентов: x, y, z,... .

n – местные предикатные константы: одноместные – “для респондента x рассматриваемый признак принимает такое-то значение”. Примеры: “возраст человека x лежит в интервале от 35 до 40 лет”; “возраст человека x лежит в интервале от 15 до 20 лет”; “профессия респондента x – врач”; “профессия респондента x – учитель” и т.д.; двуместные – “профессия респондента x не совпадает с профессией респондента y”, “респондент x читает те же газеты, что и респондент y”.

Понятие формулы определяется рекурсивно:

1) любая предикатная константа P(x), P(x,y), P (x,y,z),... является формулой;

2) если A – формула, то ¬ A – тоже формула;

3) если A и B – формулы, то A&B, A ∨ B, A ⊃ B – тоже формулы;

4) если A – формула и x – предметная переменная, то ∀xA и ∃xA – формулы;

5) ничто иное, кроме перечисленного в п.п. (1-4), формулой не является.

Будем считать, что читателю известно, как определяется истинность логических формул с кванторами всеобщности и существования (∀ и ∃) в обычной классической двузначной логике.

### *Интересующие социолога закономерности как формулы узкого исчисления предикатов*

Итак, представим себе типичную для социолога ситуацию: он осуществил опрос и перед ним лежит тысяча (может быть, не одна) анкет с ответами респондентов. Каждый ответивший характеризуется набором чисел – ответов, или, как обычно говорят, значений рассматриваемых признаков (признак соответствует вопросу).

Продолжая приведенные выше рассуждения, позволившие выразить интересующие социолога статистические закономерности (или, что для нас то же самое – результаты, получаемые с помощью известных методов анализа номинальных данных) в терминах исчисления высказываний, нетрудно придти к выводу, что более общие закономерности, в неменьшей мере важные для социолога, часто бывает возможно выразить в языке узкого исчисления предикатов. Эти закономерности означают истинность определённых формул в этом исчислении.

Приведем примеры упомянутых формул. Пусть, например, предикат (предикатная константа) P(x) означает “респондент x отметил 5-е значение 8-го признака”, предикат Q (y) - “респондент y отметил 3-е значение 14-го признака”, а предикат R(z) - “респондент z отметил 1-е значение 2-го признака. Тогда приведённое выше утверждение “ 5-е значение 8-го признака, как правило, встречается либо с 3-м значением 14-го, либо с 1-м значением 2-го” будет означать, что почти для всех x будет истинной формула $(P(x) \& (Q(x) \vee R(x)))$.

Теперь предположим, что P(x) означает “респонденту x отвечает 2-е значение 3-го признака”, Q(x) – “ респонденту отвечает 5-е значение 4-го признака, R(x) – предикат “значение 6-го признака для респондента x равно или 2, или 3”. Тогда выражение “из того, что 3-й признак принимает 2-е значение одновременно с тем, что 4-й принимает 5-е значение, как правило, следует, что 6-й признак принимает либо 2-е, либо 3-е”,” и т.д. означает, что почти для всех x будет истинно выражение $((P(x) \& Q(x)) \supset R(x))$.

Пусть S(x) – “значение 23-го признака для респондента x равно 2”, T(x) – “значение 7-го признака для респондента x равно 4”. Тогда утверждение “из того, что 23-й признак принимает какое-либо значение, кроме 2-го, следует, что 7-й признак принимает 4-е значение” будет эквивалентно утверждению истинности формулы $(\neg(S(x)) \supset T(x))$.

Нетрудно видеть, что таким образом в виде формул узкого исчисления предикатов действительно можно выразить очень многие интересующие социолога “закономерности”, “скрывающиеся” в эмпирических данных. А если учесть, что большинство методов анализа номинальных данных, как было показано в предыдущих параграфах, позволяет выявлять “закономерности” именно такого вида, то можно сказать, что практически все интересующие социолога закономерности выражаются на языке формул исчисления предикатов первого порядка.

Итак, наиболее типичной задачей, решающейся на основе анализа такого рода данных можно считать следующую: найти логическую функцию от значений признаков (выступающих в качестве предикатов), истинную для изучаемой совокупности респондентов. Получаемые выводы (найденные закономерности) могут иметь, например, такой вид (используем обычную логическую символику, логические связки соединяют записанные в неформальном виде значения рассматриваемых предикатов-признаков): "(((Проживающий в крупном городе) & (мужчина-предприниматель) & (старше 40 лет)) ∨ ((пенсионер) & (имеющий высшее экономическое образование))) ⊃ (собирается голосовать на ближайших выборах за кандидата N)".

Очевидно сходство такой постановки задачи с тем, что было обсуждено выше в п.п. 2.4.2, 2.5.3 и 2.5.4.

Теория измерений позволяет существенно повысить эффективность решения задачи поиска закономерностей описанного вида. Суть соответствующего подхода заключается в том, что упомянутые логические функции считаются аксиомами, задающими изучаемую ЭС (ей отвечает МС – фрагмент многомерного пространства). Разработаны способы внесения в определение и ЭС, и МС вероятностных характеристик. Предложены алгоритмы поиска таких аксиом. Рассмотрим соответствующий процесс более подробно**.**

***Вид искомых аксиом***

Возможность экспериментального выявления аксиом, описывающих нашу ЭС, обеспечивается тем, что необозримая совокупность всех возможных формул, подлежащих проверке, сводится к множеству, вполне поддающемуся обзору множеству (формулы этого множества служат гипотезами для проверки на ЭС). А именно, на основе положений математической логики доказываются следующие утверждения.

Совокупность формул интересующего нас характера может быть сведена к совокупности формул вида

$$C= (A_1\& \ A_2\& \ \dots \ \& \ A_к \supset A_0), \qquad (7)$$

где $A_i$ – или наши предикатные константы с произвольными предметными переменными, или их отрицания. Назовем формулы вида (7) *правилами*.

Введем также понятие *подправила* правила (7) как такой формулы, которая является импликацией, содержащей в качестве посылки – часть посылки формулы вида (1) (получающуюся за счет отбрасывания некоторых $A_i$ ), а в качестве заключения – либо то же заключение, что и в (7) (т.е. $A_0$), либо отрицание одной из тех $A_i$, (i = 1, …, k), которые не вошли в посылку. Ясно, что каждое подправило правила (7) является в то же время неким правилом того же вида (7).

Из логики и методологии науки известно, что *законами* можно считать те из гипотез, которые при одинаковой их подтвержденности на экспериментальных данных наиболее фальсифицируемы, просты и/или содержат наименьшее число параметров (ср. наше обсуждение понятия закономерности в п. 2.5.3).

Ясно, что подправило – логически **более сильное** утверждение, чем само правило. Другими словами, из истинности подправила следует истинность правила. К примеру, рассмотрим правило “из конъюнкции "быть мужчиной и жить на селе" следует "быть

курящим"” и два его подправила: (а) “из свойства "быть мужчиной" следует "быть курящим"“ и (б) “из свойства "быть мужчиной" следует "не жить на селе"“. То, что первое подправило логически более сильно, чем правило, представляется очевидным: если из свойства "быть мужчиной" следует свойство "быть курящим", то последнее следует также и из конъюнкции свойств "быть мужчиной и жить на селе". Относительно же второго подправила можно заметить, что если оно истинно, то, очевидно, конъюнкция "быть мужчиной и жить на селе" ложна. Значит, наше правило истинно в силу ложности его посылки (напомним, что, в соответствии с правилами формальной логики, из лжи следует что угодно).

Кроме того, любое подправило является и *более фальсифицируемым*, чем правило, так как содержит более слабую посылку и, следовательно, применимо к большему объему данных и тем самым в большей степени подвержено фальсификации; и *более простым*, так как содержит меньшее число атомарных высказываний, чем правило; и *включает меньшее число "параметров"*, так как лишние атомарные высказывания также можно считать параметрами "подстройки" высказывания под данные.

Обычно используемое в рамках теории измерений обоснование нефальсифицируемости какого-либо положения не предполагает поиска более простого, логически более сильного и также нефальсифицируемого утверждения. Поэтому нефальсифицируемое на имеющихся данных утверждение принимается в качестве аксиомы даже в том случае, если оно содержит некоторые дополнительные условия, которые без ущерба для нефальсифицируемости можно было бы удалить из него (скажем, мы считаем аксиомой положение "мужчины – селяне курят", если оно истинно на всех объектах изучаемой выборки, и делаем это даже тогда, когда истинным является также логически более сильное положение "мужчины курят", т.е. когда свойство "быть жителем села" – явно лишнее в аксиоме). Авторы цитируемой работы предлагают осуществлять такое удаление.

Сформулированные выше положения дают основания считать, что задача обнаружения законов в данных (законов, характеризующих изучаемую ЭСО) требует нахождения среди всех правил вида (7) логически наиболее сильных. Будем называть *законом ЭС* любое истинное на этой системе правило вида (7), для которого каждое его подправило уже не истинно на той же системе. Наша главная задача состоит в поиске таких законов, т.е. в поиске наиболее сильной теории, вытекающей из соотношений вида (7) и описывающей эти данные.

Задача вполне решаема, что подтверждается тем, что описанный подход реализован на ЭВМ [Витяев, 1992; Витяев, Москвитин, 1985, 1993]. На этом мы закончим в основном изложение базирующихся на идеях РТИ принципов поиска логических закономерностей,

характеризующих изучаемую ЭС. Сделаем лишь несколько небольших замечаний о том, чего мы пока не коснулись.

Заметим, что поиск законов может также способствовать проверке истинности на ЭС любой заранее данной системы аксиом: аксиома будет выполнена на ЭС, если найдется такое ее подправило, которое является законом. Последнее утверждение опирается на то, что, как доказано в цитируемой работе, истинность правила вида (7) возможна только в силу истинности некоторого его подправила либо первого, либо второго определенного нами вида (см. определение подправила). При этом истинность подправила второго вида имеет место в том случае, когда посылка формулы (7) ложна (напомним, что ложность посылки импликации означает истинность последней).

В рассматриваемой работе предлагается также определение вероятностного закона на изучаемой ЭС. Понятие истинности закономерности при этом заменяется на некоторую оценку ее предсказания, вероятности (что представляется целесообразным в свете описанной в первой части настоящей работы статистичности интересующих социолога законов). Рассматривается также проблема т.н. шумов – искажениями искомых законов, вызванных разными случайными причинами.

## 2.6. Анализ связей типа "признак - группа признаков": номинальный регрессионный анализ (НРА)

### 2.6.1. Общая постановка задачи

Вспомним некоторые рассуждения, использованные нами выше (п.2.2) в процессе осмысления предложенной классификации методов изучения связей между номинальными переменными. Мы подчеркивали, что в большинстве реальных задач исследователь не должен следовать ставшему традиционным ограничению круга используемых математических методов только известными коэффициентами парной связи. При этом описывалось две совокупности факторов, обусловливающих необходимость перехода к другим методам (см. рис. 20) .

Во-первых, имеет смысл "рассыпать" все рассматриваемые признаки на отдельные альтернативы и затем, "склеивая" их разными способами, искать такие сочетания значений исходных признаков, которые определяют те или иные связи, то или иное "поведение" респондентов (анализ фрагментов таблиц сопряженности, алгоритмы последовательных разбиений типа и т.д. ).

Во-вторых, имеет смысл объединять отдельные признаки друг с другом, искать такие их сочетания, которые в каком-то смысле детерминируют другие признаки и их сочетания (как мы увидим ниже, в регрессионном анализе речь пойдет о детерминации среднего уровня этих “других” признаков). К соответствующим рассмотрениям мы и перейдем в настоящем параграфе. Проанализируем ту группу методов (или задач, мы говорили о том, что задачи для нас в определенном смысле отождествляются с методами), которая при классификации задач была символически обозначена нами как методы типа "признак-(группа признаков)". Сюда относится регрессионный анализ, к рассмотрению которого мы и переходим.

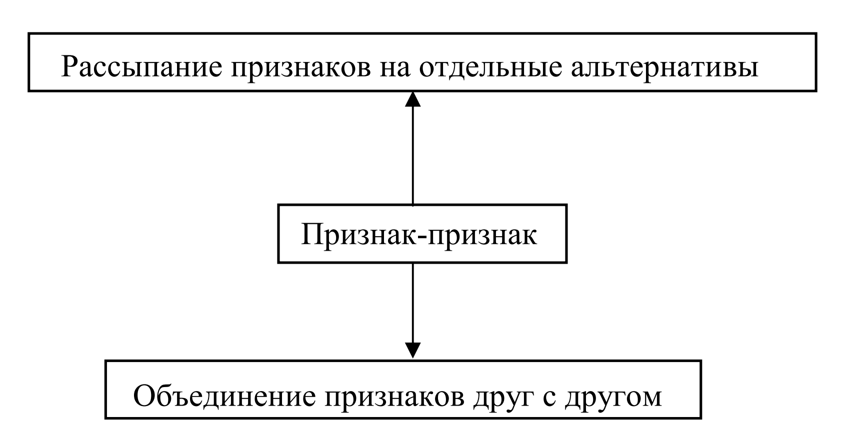



**Рис. 20. Схематичное выражение причин, обусловливающих необходимость перехода от традиционных коэффициентов парной связи к другим методам анализа связей**

Сначала для простоты изложения рассмотрим случай, когда у нас имеется только два признака – X и Y - и нас интересует зависимость между ними. Другими словами, сначала предположим, что наша "группа признаков" состоит из одного признака – X (потом перейдем к случаю, когда вместо одного X фигурируют несколько признаков). Мы знаем, что о связи между признаками говорит соответствующий коэффициент корреляции: чем ближе значение модуля этого коэффициента к 1, тем более сильна эта связь, т.е. тем с большей уверенностью мы можем полагать, что с ростом значений одного признака растут (если коэффициент корреляции положителен) или убывают (если коэффициент корреляции отрицателен) значения другого (напомним, что коэффициент корреляции измеряет линейную связь между переменными; отметим, однако, что приводимые рассуждения справедливы и для других коэффициентов связи, например, для корреляционного отношения, дающего возможность оценить криволинейную связь). Но при этом мы совершенно не можем сказать о том, в какой степени

возрастет значение Y, если значение X увеличится, скажем, на 1. А ситуации здесь могут быть весьма разными.

Приведем пример, рассмотрев зависимость между производственным стажем человека и его зарплатой. Предположим, что мы имеем дело с двумя крайними ситуациями, отраженными на рисунках 21а и 21б. В обоих случаях соответствующие коэффициенты корреляции близки к 1 (обе совокупности

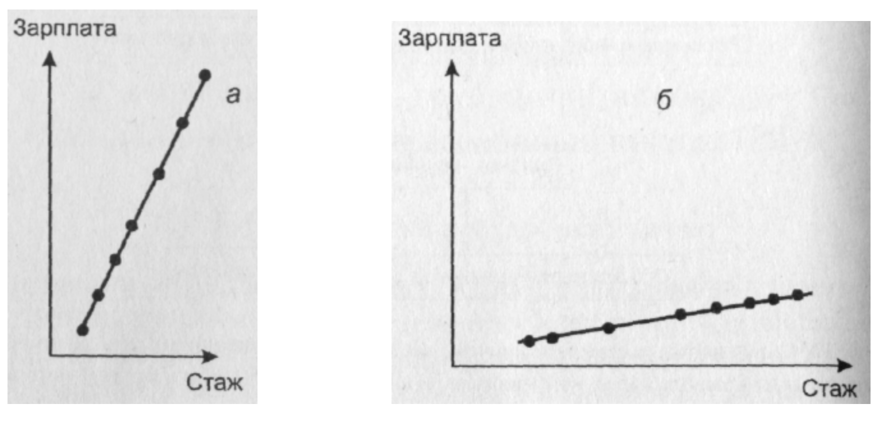



**Рис. 21. Примеры сильных линейных связей, определяющих разный прогноз**

точек-объектов лежат на прямых линиях, отвечающих нашей зависимости). На первом из них прямая идет резко вверх. Поэтому даже при небольшом увеличении X признак Y резко возрастет. В случае же наличия связи, изображенной на втором рисунке, прямая близка к горизонтали. Поэтому даже при значительном росте X значение Y почти не изменится. Другими словами, на основании наших двух картинок мы получим прогнозы совершенно различного характера. И совершенно ясно, что этого никак нельзя узнать лишь на основе вычисления соответствующих коэффициентов корреляции.

Итак, для того, чтобы делать прогноз о том, как изменится значение Y при том или ином изменении значения X, нам желательно знать, как говорят, форму связи между этими переменными, т.е. желательно найти функцию вида Y = f (X). Подчеркнем, что отношение между X и Y несимметрично: речь идет именно о зависимости второй переменной от первой, именно о возможности прогноза значения Y от X, а не наоборот.

В данном случае для обозначения X и Y используются те же термины, о которых шла речь в начале п. 2.5.3.1. Однако для той ситуации, когда речь идет о нахождении формы зависимости Y от X, употребляется еще несколько пар терминов: независимые переменные называют *входными, экзогенными, внешними*, а зависимая – *выходной, эндогенной, внутренней*. Представляется важным правильное понимание причин использования такой терминологии.

Поиск функции f предполагает разработку определенной модели связи между переменными, опирающуюся на априорные знания исследователя (так, ниже мы будем говорить в основном о линейной модели, о *линейном* регрессионном анализе). Найденная с помощью регрессионной техники зависимость – это тоже некоторая модель реальности - модель, в соответствии с которой и находятся значения Y на основе информации о значениях признака X.

Независимые признаки (X) потому и можно назвать независимыми, что они не зависят от этой модели. Эти признаки как бы поступают на ее “вход”, являются внешними по отношению к ней, берутся “со стороны”. Они определяют конкретный вид искомой зависимости, но не определяются ею. Прогнозируемые же значения зависимой переменной (Y) полностью определяются моделью (то, насколько они близки к реальности, зависит от качества модели), служат ее “выходом”, являются ее порождением. Они внутренне по отношению к ней.

Особенно осторожно надо использовать словосочетания "признак-причина" и "признак-следствие", о чем мы уже говорили в п. 2.1.3.

### 2.6.2. Повторение основных идей классического регрессионного анализа, рассчитанного на т. н. "количественные" признаки

Сначала для простоты и возможности геометрического изображения основных положений регрессионного анализа предположим, что у нас всего две переменные: X и Y (соответственно, независимая и зависимая). С помощью рассматриваемого подхода осуществляется поиск зависимости вида Y = f(X). Однако это выражение для результата регрессионного анализа носит условный характер: искомая зависимость не функциональна, а статистична, является закономерностью “в среднем”, она “неточна”. Поясним, в чем именно состоят такие усредненность и “неточность”.

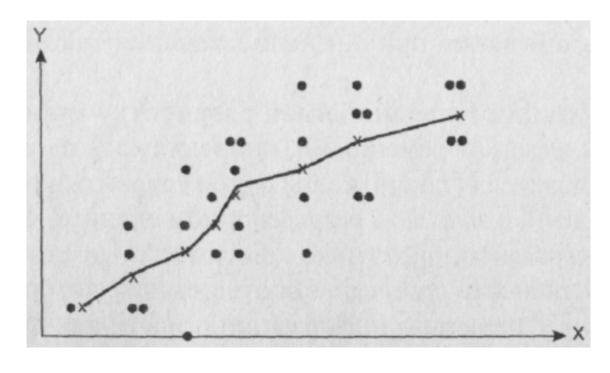



**Рис. 22. Принципиальная схема линии регрессии.**

В качестве независимой переменной фигурируют условные средние значения Y (каждое такое среднее вычисляется для конкретного значения независимой переменной X; соответствующая точка на графике обозначена крестиком)

Прежде всего обратим внимание читателя на то, что для социологических данных типична ситуация, когда одному значению X соответствует множество значений Y. Эта ситуация схематично изображена на рис. 22 (пока обращаем внимание только на черные кружки).

Встает вопрос: какую именно зависимость мы хотим вычислить? Как искомая кривая (а мы хотим, чтобы каждому значению независимой переменной отвечало одно значение зависимой, т.е. чтобы искомой связи отвечала какая-то одномерная линия) должна "пробиваться" через изображенное на рисунке облако точек?

Ответ представляется естественным: подсчитаем для каждого значения X среднее арифметическое значение всех отвечающих ему значений Y и будем изучать зависимость от X именно таких средних. Соответствующие точки на нашем рисунке обозначены крестиками. Для них вид искомой зависимости четко "просматривается". Другими словами, интересующая нас статистическая зависимость будет иметь вид:

$$\overline{Y}_X = f(X) \qquad (8)$$

Вспомним, что на рис. 22 отражена выборочная ситуация, в то время как в действительности нас интересует то, что делается в генеральной. Рассмотрение последней предполагает, что переменные непрерывны, имеют бесконечное число значений. Соотношение (8) для генеральной совокупности превращается в следующее:

$$\mu(Y / X) = f(X), \qquad (9)$$

(где μ – знак математического ожидания – меры средней тенденции для генеральной совокупности; напомним, что среднее арифметическое, является лишь "хорошей" выборочной оценкой математического ожидания). Такая функция называется *функцией регрессии* Y по X (иногда говорят об *уравнении регрессии*, либо о *регрессионной зависимости*). Ее график называется *линией регрессии*. Подчеркнем, что соотношение (9) предполагает, что при каждом фиксированном значении X значения Y суть значения некоторой случайной величины. Это означает следующее.

Фиксируя какое-либо значение X, равное, например, $X_i$ (т.е. рассматривая совокупность объектов, обладающих этим значением), мы имеем дело с некоторым условным распределением

Y (которое образуют значения зависимой переменной Y, вычисленные для объектов, обладающих значением $X_i$ признака X). Это распределение имеет свое математическое ожидание и дисперсию. Именно это математическое ожидание фигурирует в левой части равенства (9). Это математическое ожидание лежит на линии регрессии (рис. 23).

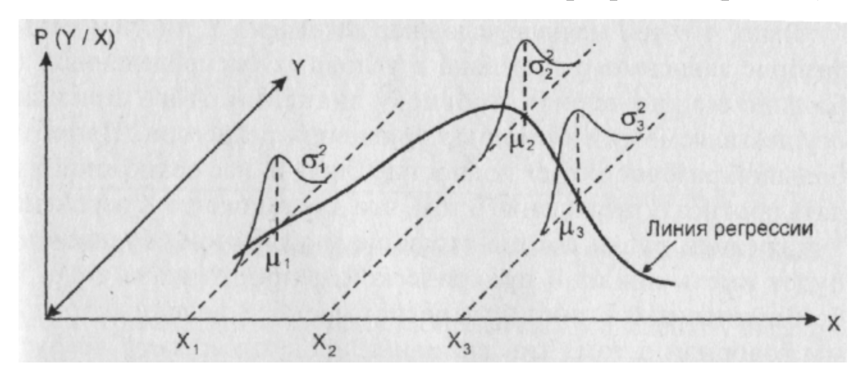



**Рис. 23. Статистические предположения, лежащие в основе регрессионного анализа.**

Условные распределения зависимой переменной Y нормальны. Их математические ожидания $\mu_1$, $\mu_2$, $\mu_3$ лежат на линии регрессии; дисперсии $\sigma_1^2$, $\sigma_2^2$, $\sigma_3^2$ равны.

$\mu_1$, $\mu_2$, $\mu_3$ – математические ожидания тех условных распределений переменной Y, которые получаются при фиксации, соответственно, значений $X_1$, $X_2$, $X_3$ переменной X. Ясно, что с помощью линии регрессии хорошо можно осуществлять тот прогноз, который является основной целью поиска зависимости Y от X: эта линия говорит о том, насколько изменится среднее значение Y при том или ином изменении значения X. Будем говорить в таком случае об изменении Y в среднем.

Точность, с которой линия регрессии Y по X передает изменение Y в среднем при изменении X, измеряется дисперсией величины Y, вычисленной для каждого значения X:

$D\,(Y/X) = \sigma^2\,(X)$

Пусть $\sigma_1^2$, $\sigma_2^2$, $\sigma_3^2$ – значения дисперсий, вычисленных для условных распределений переменной Y, получающихся при фиксации, соответственно, значений $X_1$, $X_2$, $X_3$ переменной X.

Обычно предполагается, что описанные условные распределения зависимой переменной Y нормальны, а дисперсии этих распределений равны: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$. Именно такая ситуация отражена на рис. 23. При равенстве дисперсий говорят, что условные распределения удовлетворяют свойству *гомоскедастичности*. Попытаемся коротко пояснить смысл этого свойства.

Ясно, что чем меньше условные дисперсии Y, т.е. чем меньше разброс зависимого признака в условных распределениях, тем больше можно верить прогнозу значений этого признака, осуществляемому с помощью уравнения регрессии. Напротив, большой разброс может полностью лишить нас возможности делать прогноз: утверждение о том, что для такого-то $X_i$ переменная Y в среднем равна соответствующему условному среднему, не будет иметь никакой практической ценности из-за того, что бессмысленным станет сам расчет средней величины (в п. 1.2 мы говорили о том, что для осмысленности средней требуется однородность изучаемой совокупности объектов, отсутствие большого разброса по рассматриваемому признаку). Можно говорить о качестве найденной регрессионной зависимости, связывая его именно с описанной возможностью прогноза. Тогда при условных дисперсиях, равных одной и той же величине σ, это качество может быть строго определено: при большой σ оно будет плохим, при малой – хорошим. А если разбросы при разных X разные? Тогда для одних значений X уравнение регрессии будет хорошим, при других – плохим. Представляется, что при практическом использовании такого уравнения могут возникнуть неприятности. Отсюда – требование гомоскедастичности.

Теперь обсудим вопрос о том, как найти конкретный вид функции регрессии f. На помощь приходит то, что линия регрессии обладает замечательным свойством: среди всех действительных функций f минимум математического ожидания $\mu(Y–f(X))^2$ достигается для функции f (X) = μ(Y/X). Поясним смысл этого утверждения, обратившись к выборочной ситуации, представленной на рис. 24.

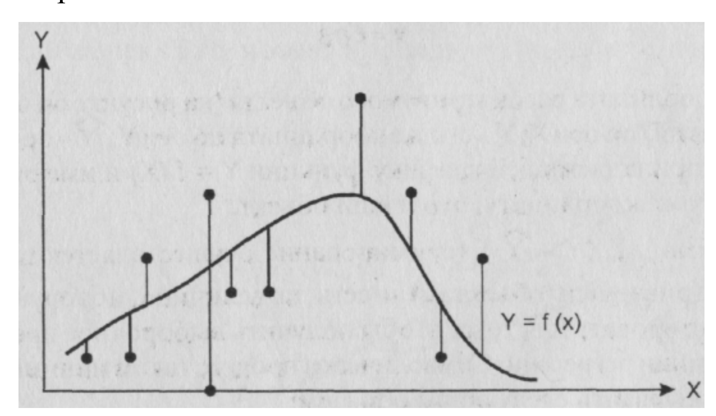



**Рис. 24. Отклонения ординат рассматриваемых точек от произвольной функции**

Рассмотрим заданную совокупность точек – моделей изучаемых объектов и произвольную функцию f (X). Вертикальные отрезки – отклонения ординат рассматриваемых

точек от этой графика этой функции. Средняя величина квадратов длин этих отрезков – это и есть выборочная оценка математического ожидания $\mu(Y-f(X))^2$.

Для того, чтобы лучше понять способ вычисления величин рассмотренных отрезков, покажем, в чем он состоит, на примере одной точки, имеющей произвольные координаты (X, Y) в нашем признаковом пространстве. Обратимся к рис. 25.

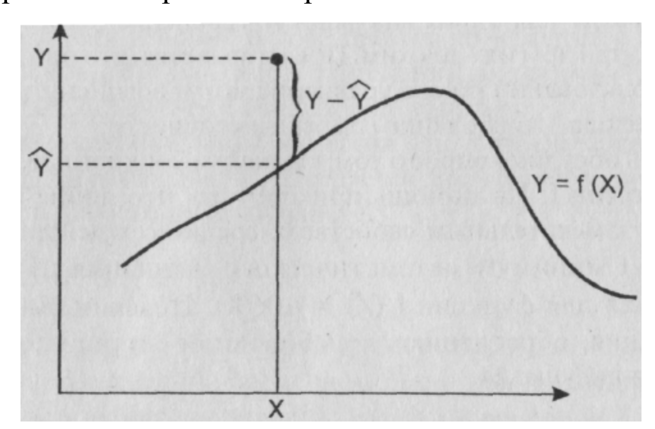



**Рис.25. Способ определения отклонения точки (X, Y) от произвольной функции Y = f (X)**

X координата рассматриваемого объекта (на рисунке он обозначен точкой) по оси X; Y – его же координата по оси Y;Y - ордината точки, принадлежащей графику функции Y = f(X) и имеющей по оси X ту же координату, что и наш объект.

Сумма $\sum(Y-\hat{Y})^2$ (суммирование осуществляется по всем рассматриваемым объектам) и есть та величина, которую надо минимизировать для того, чтобы получить выборочное представление линии регрессии. Символически процесс такой минимизации можно выразить следующим образом:

$$\sum(Y-\hat{Y})^2 \to \min \qquad (10)$$

$\hat{Y}$ – это как бы теоретическое, модельное значение зависимой переменной. Это то значение, которое мы имели бы, если бы после всех расчетов пользовались найденной функцией Y = f (X) как основой для прогноза.

В соответствии со сформулированным выше свойством линии регрессии, можно сказать, что минимальной эта сумма будет в том случае, если рассматриваемая функция Y = f(X) является выборочным представлением искомой линии регрессии. Другими словами, указанному выборочному представлению отвечает та функция f(X), для которой указанная выше сумма минимальна.

Итак, чтобы найти выборочную линию регрессии, необходимо как бы “перебрать” все возможные функции Y = f(X), для каждой вычислить указанную сумму квадратов и остановиться на той функции, для которой эта сумма минимальна.

Рассматриваемый способ поиска f(X), носит название метода наименьших квадратов (отметим, что этот метод очень часто используется при расчете самых разных статистических закономерностей. Так, он задействован в одном из известных методов шкалирования - методе парных сравнений [Толстова, 1998]).

Чтобы смысл метода наименьших квадратов стал яснее, заметим, что чем меньше величина указанной выше суммы квадратов, тем с большим основанием рассматриваемую функцию можно считать близкой одновременно ко всем рассматриваемым точкам. Эта функция в каком-то смысле служит моделью всего "облака" точек. Это можно проиллюстрировать с помощью рисунка 26.

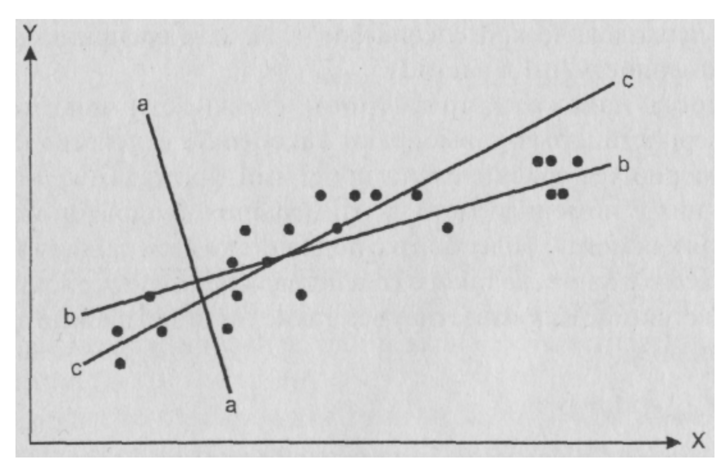



**Рис. 26. Иллюстрация проблемы выбора прямой линии, наилучшим образом отвечающей линии регрессии**

Ясно, что прямая "aa" заведомо не может минимизировать рассматриваемую сумму: она совсем не отражает наше облако точек. А вот относительно прямых "bb” и "cc" вряд ли “на глаз” можно определить, какая из них лучше. Чтобы ответить на этот вопрос, необходимо использовать метод наименьших квадратов.

Очевидно, перебрать все мыслимые функции невозможно. Встает вопрос, как определить f(X).

Математика предоставляет нам возможность найти функцию, отражающую искомую линию регрессии с любой степенью приближения. Это можно сделать, например, используя многочлены произвольной степени m:

$$g(X, \beta) = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_m X^m$$

($\beta_{0,}$ $\beta_1$, $\beta_2$, …, $\beta_m$ – некоторые параметры; выборочные оценки которых надо получить). Однако найденная функция, вообще говоря, будет очень сложной и вряд ли с ее помощью мы сможем практически осуществлять прогноз, т.е. достигнем основной цели построения регрессионных моделей. Причины такой непригодности сложных формул частично сходны с теми, что были обсуждены нами в п. 2.5.3.2 при рассмотрении третьей причины останова алгоритма THAID: слишком сложные формулы мы в силу своей психологической специфики не можем воспринимать как закономерность (п.1.4 части I).

Чтобы избежать чрезмерной сложности искомой закономерности, обычно выбирают какое-либо семейство кривых, выражающихся сравнительно простыми формулами, и именно среди них с помощью метода наименьших квадратов ищут ту, которая как можно более близко подходит ко всем данным точкам. Чаще всего в качестве такого семейства используют совокупность прямых линий. Как известно, все такие линии выражаются формулами вида

$$g(X, \beta) = \beta_0 + \beta_1 X$$

где $\beta_1$ а говорит о величине угла наклона прямой к оси X, а $\beta_0$ - о сдвиге этой прямой вдоль оси Y. Соответствующий вариант регрессионного анализа называется *линейным*. Он чаще всего используется практически. Отвечающая ему техника хорошо известна. Выборочные оценки коэффициентов линейного уравнения регрессии находятся с помощью описанного выше метода наименьших квадратов.

В данном случае (10) превращается в соотношение

$$\sum (Y - \beta_0 + \beta_1 X)^2 \to \min$$

Далее мы, условно говоря, как бы “перебираем” все возможные прямые (точнее, все возможные пары чисел $\beta_0$ и $\beta_1$) и находим ту прямую, для которой наша сумма будет самой маленькой. Конечно, в действительности перебрать все прямые также невозможно (как известно, совокупность всех действительных чисел нельзя даже “пересчитать” с помощью бесконечного ряда натуральных чисел), параметры искомой прямой ищутся с помощью производных: находим производную от нашей суммы по $\beta_0$ и $\beta_1$ и ищем те их значения, которые обращают производную в нуль. Получаем известные аналитические выражения для этих коэффициентов (напомним, что латинскими буквами обозначаются выборочные оценки одноименных генеральных параметров):

$$b_0 = \overline{Y} - b_1 \overline{X} = r \frac{S_Y}{S_X}$$

$$b_1 = \sum \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{(X_i - \overline{X})^2}$$

где r – коэффициент корреляции между X и Y; $S_Y$ и $S_X$ – выборочные оценки средних квадратических отклонений соответствующих признаков; суммирование, как и выше, осуществляется по всем объектам.

В идеале точка с координатами (X, $\beta_0 + \beta_1 X$) должна лежать на линии регрессии. В соответствии с упомянутыми выше традиционными предположениями, это означает справедливость картины, отраженной на рис. 27.

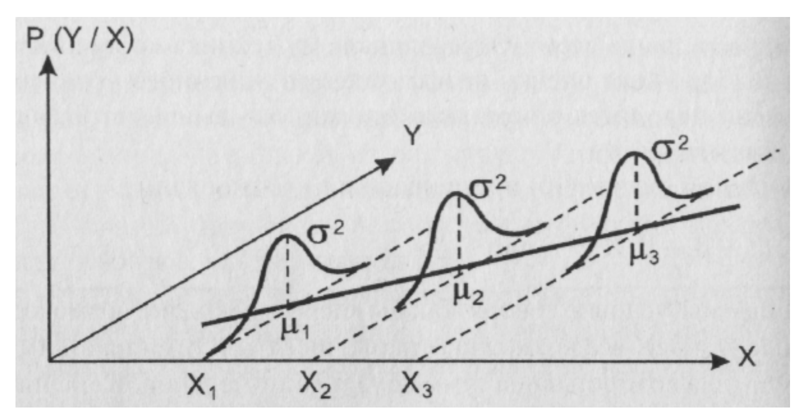



**Рис. 27. Статистические предположения, лежащие в основе линейного регрессионного анализа.**

Условные распределения Y нормальны. Их математические ожидания лежат на прямой линии, дисперсии равны.

Другими словами, мы предполагаем, что каждому значению независимой переменной X отвечают нормальные гомоскедастичные условные распределения Y, математические ожидания которых принадлежат рассматриваемой прямой. Это предположение эквивалентно следующему соотношению:

$Y_i = \beta_0 + \beta_1 X_i + e_i$,

означающему, что каждое наблюдаемое значение $Y_i$ есть сумма некой фиксированной величины $\beta_0 + \beta_1 X$, обусловленной линией регрессии, и случайной величины $e_i$, обусловленной естественной вариацией значений Y вокруг линии регрессии. При каждом значении независимой переменной X вариация Y имеет тот же характер, что и вариация $e_i$. Отсюда ясно, что все $e_i$ имеют нормальные распределения с нулевыми математическими ожиданиями и равными дисперсиями $\sigma_2$. Важность случайных величин $e_i$ заключается в том, что она

представляет собой главный источник ошибок при попытке предсказать Y по значению X. В рамках регрессионного анализа разработаны способы оценки величин $e_i$.

На практике чаще всего пользуются именно линейными регрессионными моделями. Однако при их использовании необходимо учитывать, что идеальная картина, изображенная на рис. 27 – это лишь наше пожелание. Наилучшая прямая среди всех возможных прямых может быть весьма плохим приближением к реальности. Скажем, если наши крестики расположены так, как это отражено на рис. 28, то любая прямая (например, "аа") здесь даст очень плохое приближение.

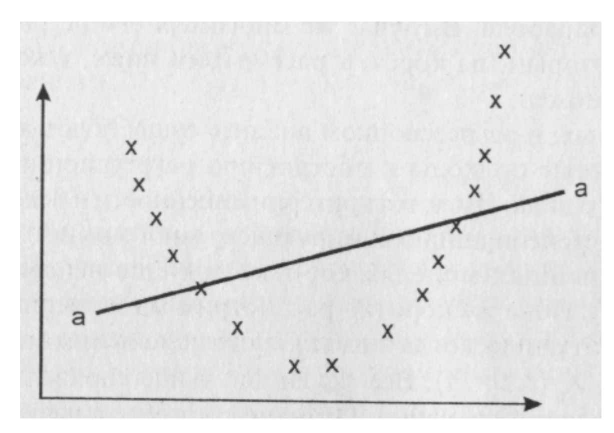



**Рис. 28. Пример криволинейной линии регрессии между двумя переменными.**

Несоответствие ей прямой "аа"

В данном случае надо бы вместо прямых линий использовать для поиска подходящих кривых семейство квадратных трехчленов вида

$Y = \beta_0 + \beta_1 X + \beta_2 X_2$.

Используя же технику линейного регрессионного анализа, и тем самым направляя свою энергию на поиск лучшей прямой, приближающей нашу совокупность точек, мы рискуем никогда не узнать, что в действительности имели дело с линией регрессии, являющейся параболой. Правда, тут необходимо отметить два момента.

Во-первых, для двумерного случая, который мы пока рассматриваем, такое вряд ли случится, поскольку перед нами – наглядная плоскостная картина, глядя на которую всегда можно определить, прямая ли линия соответствует изучаемому множеству точек, или парабола. В случае же многомерного регрессионного анализа, который мы коротко рассмотрим ниже, такой просчет вполне возможен.

Во-вторых, в регрессионном анализе существуют достаточно разработанные подходы к построению регрессионных кривых нелинейного вида. Имеются критерии линейности и рекомендации по выбору степени аппроксимирующего многочлена.

О нелинейных моделях коротко мы еще вспомним ниже (см. п. 2.6.5). Пока же коротко рассмотрим многомерный случай, т.е. такую ситуацию, когда имеется много независимых переменных $X_1$, $X_2$, ..., $X_n$ ($n > 1$). Все сказанное выше справедливо и для рассматриваемой ситуации. Отличие состоит только в том, что здесь линейная регрессионная модель имеет вид не прямой линии, а так называемой гиперплоскости:

$$Y = a_0 + a_1 \times X_1 + a_2 \times X_2 + ... + a_n \times X_n$$

Здесь необходимо два слова сказать об интерпретации только что выписанного уравнения (в соответствии с общепринятой терминологией, слева пишется просто Y, а не условное среднее $\overline{Y}_{X_1 X_1 \ldots X_n}$ и найденное с помощью техники регрессионного анализа соотношение называется уравнением, хотя этот термин и употребляется не в том смысле, в каком его используют в школе; $a_0$ называется свободным членом уравнения). Однако прежде сделаем некоторые замечания о единицах измерения рассматриваемых признаков. Интуитивно ясно, что уравнение регрессии будет более ясным с точки зрения его содержательной интерпретации, если все эти единицы будут одинаковыми. Для этого обычно осуществляют так называемую стандартизацию всех значений каждого признака: вычитают из каждого такого значения среднее арифметическое признака (точнее, здесь речь должна идти о математическом ожидании, за неимением которого мы используем его выборочную оценку – среднее арифметическое) и делят полученную разность на его же дисперсию (и снова вместо генеральной дисперсии мы вынуждены пользоваться ее выборочной оценкой). Рассмотрим для примера признак $X_2$.. Если $X_2^i$ – некоторое (i-е) его значение, $\overline{X}_2$ и $\sigma_X$ – соответственно, отвечающие ему среднее арифметическое и дисперсия, то указанная нормировка будет означать следующее преобразование значения $X_2^i$:

$$X_2^i \to \frac{X_2^i - \overline{X}_2}{\sigma_X}$$

Нетрудно видеть, что среднее значение нормированного признака будут равно нулю, а дисперсия – единице. Далее будем считать, что описанная нормировка для всех

рассматриваемых признаков произведена и что тем самым снята проблема несравнимости их значений из-за “разномасштабности”. Обозначения признаков оставим прежними.

Интерпретация коэффициентов очевидна. Если, скажем, значение признака $X_2$ изменится на единицу, то значение Y изменится на $a_2$. Поэтому $a_2$ можно интерпретировать как величину приращения Y, получаемого за счет увеличения признака $X_2$ на единицу.

В заключение обсуждения вопроса о классическом регрессионном анализе заметим, что указанная “прозрачная” интерпретация может “затуманиться” в том случае, если наши предикторы связаны друг с другом. Причина тоже довольно очевидна. Поясним это.

Предположим, что $X_2$ связан с $X_5$ и мы хотим узнать, на сколько изменится Y при увеличении $X_2$ на единицу. Рассуждать так же, как выше, мы не можем: увеличение $X_2$ неумолимо приведет к увеличению (или уменьшению) $X_5$, и поэтому изменение Y будет обусловлено изменением не только $X_2$, но и $X_5$. На сколько изменится $X_5$, вообще говоря, неизвестно. Чтобы ответить на этот вопрос, нужно подробнее изучить форму зависимости между $X_2$ и $X_5$. А это - самостоятельная и, возможно, сложная задача. Без ее решения вопрос о величине изменения Y остается открытым. И в любом случае это изменение, вообще говоря, не будет равно $a_2$.

В силу сказанного, будем стремиться к тому, чтобы избегать включения в уравнение регрессии заведомо связанных друг с другом предикторов.

Описание идей регрессионного анализа можно найти в [Мостеллер, Тьюки, 1982; Паниотто, Максименко, 1982; Статистические методы …, 1979].

Теперь перейдем к рассмотрению вопроса о возможности использования техники линейного регрессионного анализа к номинальным данным.

### 2.6.3. Дихотомизация номинальных данных. Обоснование допустимости применения к полученным дихотомическим данным любых "количественных" методов

Конечно, использовать регрессионную технику для анализа “чисел”, являются метками, отвечающими некоторой номинальной шкале, бессмысленно (считаем это интуитивно ясным, хотя можно было бы доказать такое утверждение строго, используя понятие адекватности математического метода из теории измерений (см., например, (Толстова, 1998)). Для того, чтобы на основе информации, полученной по номинальной шкале, можно было построить уравнение регрессии, эту информацию необходимо преобразовать. Соответствующее преобразование

носит название дихотомизации номинальных данных. Этот подход применяется очень широко, поскольку его использование как бы “открывает дверь” для применение подавляющего большинства “количественных” методов с целью анализа номинальных данных. Опишем суть преобразования.

Вместо каждого номинального признака, принимающего к значений, вводим к новых дихотомических (т.е. принимающих два значения, будем обозначать эти значения 0 и 1). Надеемся, что то, как это делается, станет ясным из следующего примера.

Предположим, что рассматриваемый номинальный признак X – это национальность и что в соответствующем закрытом вопросе анкеты фигурируют три национальности: русский, грузин и чукча. Каждой из этих альтернатив поставим свой дихотомический признак, задаваемый следующим правилом (напомним, что задать признак - значит задать правило приписывания отвечающих ему значений каждому респонденту):

$$\textit{русский} \rightarrow X_1 = \begin{cases} 1, \ \textit{если} \quad \textit{рассматриваемый} \quad \textit{респондент} - \textit{русский} \\ 0, \ \textit{если} \quad \textit{рассматриваемый} \quad \textit{респондент} - \textit{не} \quad \textit{русский} \\ (\textit{кто} \quad \textit{именно} - \textit{грузин} \quad \textit{или} \quad \textit{чукча} - \textit{безралично}) \end{cases}$$

$$\textit{грузин} \rightarrow X_2 = \begin{cases} 1, \ \textit{если} \quad \textit{рассматриваемый} \quad \textit{респондент} - \textit{грузин} \\ 0, \ \textit{если} \quad \textit{рассматриваемый} \quad \textit{респондент} - \textit{не} \quad \textit{грузин} \\ (\textit{кто} \quad \textit{именно} - \textit{русский} \quad \textit{или} \quad \textit{чукча} - \textit{безралично}) \end{cases}$$

$$\textit{чукча} \rightarrow X_1 = \begin{cases} 1, \ \textit{если} \quad \textit{рассматриваемый} \quad \textit{респондент} - \textit{чукча} \\ 0, \ \textit{если} \quad \textit{рассматриваемый} \quad \textit{респондент} - \textit{не} \quad \textit{чукча} \\ (\textit{кто} \quad \textit{именно} - \textit{русский} \quad \textit{или} \quad \textit{грузин} - \textit{безралично}) \end{cases}$$

Применение регрессионной техники к преобразованным номинальным данным называется номинальным регрессионным анализом. Поясним подробнее, что именно при реализации соответствующего подхода происходит с зависимой и независимыми переменными. Предположим, что мы хотим изучить связь вида

Y = f(X),

где X – скажем, та же национальность (предусматривающая, как и выше, три варианта ответов), а Y – профессия. Вместо признака X в уравнение необходимо вставить три новых предиктора – $X_1$, $X_2$, $X_3$, описанные выше. Однако здесь имеется один нюанс. В конце п. 2.6.1. мы отмечали нежелательность включения в регрессионную модель таких предикторов, которые заведомо связаны друг с другом. А относительно наших $X_1$, $X_2$, $X_3$ такая связь как раз имеет место. Покажем это.

Нетрудно видеть, что если мы знаем значения двух из трех рассматриваемых предикторов, то значение третьего определяется автоматически. Мы можем не спрашивать респондента, какая у него национальность, а сами определить ее, если знаем, какие значения для него имеют признаки $X_1$ и $X_2$. Это демонстрируется приведенной ниже таблицей 28.

*Таблица 28.*

**Иллюстрация зависимости друг от друга признаков, являющихся результатом дихотомизации одной номинальной переменной**

| Заданные значения признаков | | Теоретически определяемое значение признака |
|---|---|---|
| $X_1$ | $X_2$ | $X_3$ |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |

(если человек – не русский и не грузин, то он – чукча; если он русский, а не грузин, то он и не чукча; если же он не русский, но грузин, то он тоже не чукча; быть же одновременно и русским, и грузином он не может).

Поэтому во избежание недоразумений, могущих возникнуть при интерпретации результатов регрессионного анализа, желательно не включать в уравнение все три дихотомические переменные. Именно так обычно и поступают. Один дихотомический признак как бы отбрасывают (ниже мы увидим, что это отбрасывание в содержательном плане является фиктивным: в процессе интерпретации коэффициентов найденного уравнения сведения об отброшенном признаке будут присутствовать). Таким образом, число аргументов искомого уравнения будет на единицу меньше, чем число альтернатив в рассматриваемом номинальном признаке. В нашем случае вместо трех предикторов мы включаем в уравнение только два. Ниже будем считать, что мы отбросили $X_3$.

Теперь рассмотрим ситуацию с зависимой переменной Y. Она так же, как и X превращается в несколько дихотомических признаков. Пусть, например, в нашей анкете предусмотрено три варианта ответа - учитель, торговец, дворник. Тогда вместо Y возникают три следующие дихотомические признака:

$$Y_1 = \begin{cases} 1, & \textit{если} \quad \textit{респондент} - \textit{учитель}, \\ 0, & \textit{если} \quad \textit{респондент} - \textit{не} \quad \textit{учитель}; \end{cases}$$

$$Y_2 = \begin{cases} 1, & \textit{если} \quad \textit{респондент} - \textit{торговец}, \\ 0, & \textit{если} \quad \textit{респондент} - \textit{не} \quad \textit{торговец}; \end{cases}$$

$$Y_3 = \begin{cases} 1, & \textit{если} \quad \textit{респондент} - \textit{дворник}, \\ 0, & \textit{если} \quad \textit{респондент} - \textit{не} \quad \textit{дворник}; \end{cases}$$

Встает вопрос: какой из этих новых Y-ков необходимо взять в качестве независимой переменной искомого уравнения регрессии (ясно, что использование сразу нескольких зависимых переменных бессмысленно). Выход довольно очевиден: надо строить три уравнения регрессии, каждое из которых отвечает своему $Y_{i}$.

Итак, задача сводится к построению следующей системы уравнений регрессии (термин “система” здесь употреблен не случайно: уравнения взаимосвязаны и содержательно дополняют друг друга):

$Y_1 = f_1(X_1, X_2)$,

$Y_2 = f_2(X_1, X_2)$,

$Y_3 = f_3(X_1, X_2)$,

Как мы уже отмечали, техника нахождения конкретного вида каждого уравнения традиционна - это техника “числового” регрессионного анализа.

Попытаемся ответить на вопрос о том, почему такая подмена возможна, т.е. почему к числам, полученным по произвольной номинальной шкале, применять регрессионную технику (равно как и любой другой “количественный” метод) нельзя, а к отвечающим номинальной же шкале 0 и 1 – можно (и это “разрешение” тоже касается не только регрессионного анализа). Напомним, что аналогичный вопрос применительно к вычислению среднего арифметического

уже рассматривался нами в п.1.2. В настоящем и следующем параграфе мы обсудим его в более общей постановке.

Во-первых, с формальной точки зрения упомянутую дихотомическую номинальную шкалу можно рассматривать как частный случай интервальной. Здесь мы имеем дело только с одним интервалом – между 0 и 1. И представляется вполне допустимой истинность утверждения: за равными числовыми интервалами стоят некоторые реальные равные эмпирические разности между объектами.

Во-вторых, допустимость применения количественного метода к дихотомическим данным опирается на то, что, как оказывается, многие известные математические статистики, будучи вычисленными для таких данных, как правило, оказывается возможным проинтерпретировать вполне разумным образом, чего отнюдь нельзя сказать об интерпретации соответствующих показателей, вычисленных для многозначных номинальных шкал.

Пример вычисления среднего арифметического для пола респондента, приведенный в разделе 1, подтверждает это (отметим, однако, что полу отвечает естественная дихотомия, а не искусственная, как в рассмотренных выше ситуациях; иногда естественные и искусственные дихотомии противопоставляют друг другу; однако для нас это не актуально). Демонстрация того, что осмысленная интерпретация возможна и для найденных рассматриваемым образом коэффициентов уравнения регрессии, будет осуществлена в п. 2.6.4.

Последнее обстоятельство, на котором нам хотелось бы остановиться в данном параграфе, состоит в том, что, как оказывается, задача применения традиционной регрессионной техники остается осмысленной и для того случая, когда Y измеряется по интервальной шкале. Специфика такой ситуации проявляется в интерпретации результатов регрессионного анализа. Ниже на этом мы также остановимся.

### 2.6.4. Общий вид линейных регрессионных уравнений с номинальными переменными. Их интерпретация

Итак, предположим, что у нас имеется некоторые номинальные признаки Y (зависимый; пока, до обсуждения некоторых вопросов, связанных с интерпретацией результатов регрессионного анализа, будем считать этот признак номинальным) и $X_1$, $X_2$, ..., $X_n$ (независимые). Пусть Y принимает k значений, а каждый признак $X_i$ - $l_i$ значений. Предположим также, что осуществлена дихотомизация исходных данных, в результате чего независимый признак “превращен” в дихотомические признаки $Y_1$, $Y_2$, ..., $Y_k$, а каждый признак $X_i$ - в

дихотомические $X_1^i$, $X_2^i$, ..., $X_{li}^i$. Будем полагать, что в качестве “отбрасываемого признака фигурирует последний признак из каждого только что приведенного набора. Применение техники номинального регрессионного анализа к такого рода данным означат расчет k уравнений вида:

$$Y_1 = f_1(X_1, X_2, ..., X_n) =$$

$$= f_1(X_1^1, X_2^1, ..., X_{l_1-1}^1, X_1^2, X_2^2, ..., X_{l_2-1}^2, ..., X_1^n, X_2^n, ..., X_{l_n-1}^n)$$

$$Y_2 = f_2(X_1, X_2, ..., X_n) = f_2(X_1^1, X_2^1, ..., X_{l_1-1}^1, X_1^2, X_2^2, ..., X_{l_2-1}^2, ..., X_1^n, X_2^n, ..., X_{l_n-1}^n)$$

$$Y_k = f_k(X_1, X_2, ..., X_n) = f_k(\underbrace{X_1^1, X_2^1, ..., X_{l_1-1}^1}_{\text{отвечают } X^1}, \underbrace{X_1^2, X_2^2, ..., X_{l_2-1}^2}_{\text{отвечают } X^2}, ..., \underbrace{X_1^n, X_2^n, ..., X_{l_n-1}^n}_{\text{отвечают } X^n})$$

Не хотим далее “мучить” читателя индексами и поэтому все дальнейшие рассуждения будем вести в предположении, что рассматривается только одна градация зависимого признака с отвечающей ей дихотомической переменной Y и один принимающий три значения независимый признак X с отвечающими ему дихотомическими переменными $X_1$, $X_2$, $X_3$. Надеемся, что необходимые обобщения читатель сделает самостоятельно.

Таким образом, будем полагать, что искомая зависимость имеет вид:

$$Y = f(X_1, X_2) = a_0 + a_1 \times X_1 + a_2 \times X_2 \qquad (11)$$

Например, предположим, что Y, $X_1$, $X_2$ – это дихотомические переменные, отвечающие, соответственно, свойствам “быть торговцем”, “быть русским” и “быть грузином” (напомним, что дихотомическую переменную, отвечающую свойству “быть чукчей”, мы при построении уравнения отбрасываем). Процесс поиска подобной зависимости состоит в реализации техники линейного регрессионного анализа.

Коэффициенты уравнения регрессии, найденные по всем правилам классического регрессионного анализа, выражаются довольно сложными формулами, включающими в себя такие (вроде бы "запретные" для номинальных данных) статистики, как среднее арифметическое, дисперсия, частные коэффициенты корреляции и т.д., Однако, как мы уже упоминали, их оказывается возможным проинтерпретировать вполне разумным, понятным любому социологу, способом – как некоторые условные частоты. Опишем эту интерпретацию.

Сначала проинтерпретируем коэффициент $a_0$ (свободный член уравнения (5)). В силу самой сути уравнения регрессии, подставив в него произвольные значения независимых переменных $X_1$, $X_2$, слева от знака равенства мы получим среднее значение Y, которое отвечает

совокупности респондентов с рассматриваемыми значениями предикторов. Рассмотрим только тех людей, которым соответствует отброшенная нами национальность, – чукчей. Ясно, что для них $X_1 = X_2 = 0$. Подставив эти значения в уравнение регрессии, получим соотношение

$Y = a_0$

Таким образом, интерпретируемый коэффициент $a_0$ равен среднему арифметическому значению зависимой переменной для отброшенной категории респондентов, в данном случае – для чукчей. Если бы Y был интервальной переменной, то тем самым интерпретация свободного члена уравнения регрессии была бы окончена. Но наш Y – дихотомическая переменная, отвечающая свойству “быть торговцем”. В соответствии с описанной выше интерпретацией среднего арифметического значения дихотомического признака, смысл $a_0$ сводится к тому. Что это - доля чукчей, работающих торговцами (говоря формально – доля отброшенной категории респондентов, обладающих единичным значением зависимого признака).

Перейдем к интерпретации коэффициента $a_1$ из уравнения (11). Рассмотрим только русских. Нетрудно видеть, что для них $X_1 = 1$ и $X_2 = 0$. Подставим эти значения в уравнение. Получим соотношение:

$Y = a_0 + a_1$.

Учитывая осуществленную выше интерпретацию свободного члена уравнения, применительно к нашему примеру, можно сказать, что $a_1$ – это тот “довесок”, который надо прибавить к доле чукчей, являющихся торговцами, чтобы получить долю русских, занимающихся этим делом. Аналогична интерпретация $a_2$: это та величина, которую надо прибавить к доле торговцев среди чукчей, чтобы получить аналогичную долю среди грузин. Приведем пример.

Пусть уравнение, найденное с помощью линейного регрессионного анализа имеет вид:

$$Y = 0{,}3 - 0{,}1\,X_1 + 0{,}6\,X_2 \qquad (12)$$

Его коэффициенты можно интерпретировать как условные частоты: доля торговцев среди чукчей равна 0,3, среди русских – (0,3 + (- 0,1)) = 0,2, а среди грузин – (0.3 + 0,6) = 0,9.

Чтобы еще более стал ясен смысл коэффициентов уравнения регрессии, рассмотрим, во что это уравнение превращается в случае изучения двух дихотомических признаков. Приведем пример из [Типология и классификация ..., 1982. С. 260 - 266]. Пусть X - семейное положение (два значения: $X_1$ – женат, $X_2$ – неженат), Y – посещение кинотеатра ($Y_1$ – посещает, $Y_2$ – не посещает; здесь мы отвлекаемся от точного смысла этих слов: означает ли выражение “не посещает” то, что респондент никогда не ходил в кино, или же что он не был там в течение последних 5-ти лет и т.д.).

Пусть таблица сопряженности, отвечающая нашим признакам, имеет вид:

*Таблица 29.*

**Общий вид четырехклеточной таблицы сопряженности**

| Y | X | | Итого |
|---|---|---|---|
| | $X_1$ | $X_2$ | |
| $Y_1$ | a | b | a+b |
| $Y_2$ | c | d | c+d |
| Итого | a+c | b+d | a+b+c+d |

Найдем коэффициенты уравнения регрессии вида

Y = α + βX.

В соответствии с нашими правилами, они равны:

$\alpha = \frac{b}{b+d}$ (доля посещающих кинотеатр среди неженатых);

$\beta = \frac{a}{a+c} - \frac{b}{b+d}$ (тот “довесок”, который надо прибавить к доле посещающих кинотеатр среди неженатых, чтобы получить аналогичную долю среди женатых (последняя равна $\frac{a}{a+c}$).

Приведем соответствующий цифровой пример. Пусть конкретная матрица имеет вид:

*Таблица 30.*

**Пример четырехклеточной таблицы сопряженности**

| Y | X | | Итого |
|---|---|---|---|
| | $X_1$ | $X_2$ | |
| $Y_1$ | 48 | 38 | 86 |
| $Y_2$ | 2 | 12 | 14 |
| Итого | 50 | 50 | 100 |

Тогда верны соотношения:

$\alpha = \frac{b}{b+d} = \frac{38}{50} = 0{,}76$; $\beta = \frac{a}{a+c} - \frac{b}{b+d} = 0{,}96 - 0{,}76 = 0{,}2$

Y = 0,76 + 0,2X.

Нетрудно увидеть связь между номинальным регрессионным и детерминационным анализом. Действительно, в соответствии с последним, $I(X_2 \to Y_1) = P(Y_1/X_2) = \frac{38}{50} = \alpha$. В то же время $I(X_1 \to Y_1) = P(Y_1/X_1) = \frac{48}{50}$ и поэтому $\beta = I(X_1 \to Y_1) - I(X_2 \to Y_1)$.

Итак, все коэффициенты рассматриваемого уравнения регрессии интерпретируются через некоторые условные частоты. Встает вопрос: надо ли использовать сложную технику регрессионного анализа для того, чтобы получить результаты, получаемые обычно социологом более простым путем (путем прямого расчета многомерных частотных таблиц) ? Покажем, что такая постановка вопроса неправомерна: регрессионный анализ нельзя свести только к получению условных частот. Уравнение регрессии представляет собой систему, свойства которой не сводятся к свойствам отдельных составляющих ее элементов (коэффициентов найденного уравнения). Рассмотрим это обстоятельство подробнее.

### 2.6.5. Типы задач, решаемых с помощью НРА. Краткие сведения о логит- и пробит-моделях регрессионного анализа

Итак, *первый тип* решающихся с помощью НРА задач – это нахождение определенных условных процентов. Однако, как мы уже заметили, интерпретация результатов регрессионного анализа не сводится к интерпретации отдельных коэффициентов уравнения регрессии. Выше, в начале нашего рассмотрения этого подхода, мы говорили о том, что основная цель его использования в любой науке состоит в получении возможности определенного рода прогноза. Попытаемся проинтерпретировать модели номинального регрессионного анализа с соответствующей точки зрения.

Вернемся к модели общего вида:

$Y_1 = f_1 (X_1, X_2, ..., X_n) =$

$$= f_1 ( X_1^1, X_2^1, ..., X_{l_1-1}^1, X_1^2, X_2^2, ..., X_{l_2-1}^2, ..., X_1^n, X_2^n, ..., X_{l_n-1}^n )$$

Сначала предположим, что мы используем линейные модели.

По тому, какие из коэффициентов уравнения регрессии принимают наибольшие значения, можно судить о тех сочетаниях значений независимых признаков, которые в наибольшей мере детерминируют наличие у респондентов единичного значения зависимого. Другими словами, можно осуществлять поиск взаимодействий. Здесь явно просматривается связь с теми задачами, на решение которых направлены рассмотренные выше алгоритмы типа

AID (напомним, более или менее подробно мы рассмотрели алгоритмы THAID и CHAID в п. 2.5.3.2 и 2.5.3.3 соответственно). Это – *второй тип* задач. Опишем способы их решения более подробно.

Пусть $X_1$ – как выше, национальность с градациями (русский, грузин, чукча), $X_2$ – место проживания с градациями (город, село, кочевье), Y – дихотомическая переменная, отвечающая профессии "торговец". И если при подсчете уравнения линейной номинальной регрессии, к примеру, окажется, что сравнительно большими являются коэффициенты при дихотомических переменных $X_2^1$ (отвечающей свойству "быть грузином") и $X_1^2$ (жить в городе), то это будет означать, что именно эти два свойства в совокупности определяют тот или иной уровень доли торговцев в изучаемой группе респондентов. Представляется очевидным сходство этих выводов с теми, которые позволяют получать алгоритмы THAID и CHAID.

Еще более надежными станут выводы подобного рода, если мы будем использовать нелинейные модели. Сразу подчеркнем, что в номинальном регрессионном анализе гораздо легче решается проблема выбора модели, чем в "числовом" варианте этого анализа. Так, здесь резко сокращается круг тех многочленов, среди которых имеет смысл искать интересующие нас закономерности. В частности, ни к чему вставлять в искомое уравнение степени рассматриваемых переменных, поскольку для любого дихотомического признака любая его степень равна самому признаку (так как $0^2 = 0$, $1^2 = 1$). А вот произведения переменных имеет смысл включить. Эти произведения отвечают тем самым взаимодействиям, о которых шла речь выше.

Например, если доля торговцев среди изучаемых респондентов определяется долей горожан-грузин, то мы, несомненно, это выявим путем включения в уравнения произведения вида $X_2^1 \times X_1^2$ (обозначения – как выше).

Ясно, что произведения трех дихотомических переменных будут отвечать "трехмерным" взаимодействиям и т.д.

*Третий тип* задач связан с возможностью осуществлять прогноз несколько иного вида. Поясним это на примере. Вернемся к соотношению (12). В силу его очевидных арифметических свойств, можно сказать, что коэффициенты –0,1 и 0,6 означают вклад, соответственно, свойств "быть русским" ($X_1$) и "быть грузином" ($X_2$) в долю торговцев (Y) среди респондентов изучаемой совокупности. Однако проинтерпретировать смысл этого вклада трудно при дихотомических переменных. Поэтому часто прибегают к следующим рассуждениям, опирающимся на довольно сильные модельные предположения. Полагают, что указанное

уравнение справедливо не только для того случая, когда $X_1$ и $X_2$ – дихотомические переменные, характеризующие отдельных респондентов, но для такой ситуации, когда в качестве единиц наблюдения фигурируют группы людей, а $X_1$ и $X_2$ – доли, соответственно, русских и грузин в этих группах. В таком случае смысл уравнения становится ясным: если доля русских увеличивается в группе, скажем, на 10%, то доля торговцев увеличивается на (–0,1)×10% =–1% (т.е. уменьшается на 1%). Если же доля грузин в совокупности увеличивается на 10%, то доля торговцев увеличивается на (0,6)×10 % = 6%.

Заметим, что класс решаемых с помощью техники номинального регрессионного анализа задач может быть расширен за счет использования приемов, широко применяющихся во всем мире при анализе статистического материала, но не рассмотренных в настоящем учебнике. Мы имеем в виду т.н. обобщенные линейные модели (generalized linear model, GLM), в частности, логистическую регрессию, использование т.н. логит-моделей. Коротко опишем суть подхода, уделив особое внимание тому случаю, когда Y – дихотомическая номинальная переменная. То, о чем пойдет речь, можно найти в работах [Agresti, 1996. Ch.4; Demaris, 1992. Ch.4; Menard, 1995].

Напомним, что линейное регрессионное уравнение чаще всего имеет следующий вид:

$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$.

Левая часть этого уравнения обычно связывается со случайной компонентой рассматриваемой линейной модели. Эта компонента говорит о том, что объясняемая переменная Y является случайной величиной с математическим ожиданием μ. О правой части говорят как о систематической компоненте линейной модели. При этом понятие линейности зачастую расширяется: допускается, что одни $x_i$ могут выражаться через другие. Например, наличие переменной вида $x_3 = x_1 x_2$ говорит о взаимодействии между $x_1$ и $x_2$ в процессе их воздействия на Y. Наличие переменной вида $x_3 = x_1^2$ свидетельствует о криволинейности воздействия $x_1$ на Y.

Очень важным элементом рассматриваемой модели является форма связи между случайной и систематической компонентами модели. Выше мы говорили о сложности выбора этой формы. Но при этом полагали, что разные виды зависимости можно отразить с помощью преобразования правой части модели. Однако имеет смысл преобразовывать и левую часть. Так, в литературе по анализу данных принято называть *связующей функцией* (link function) такую функцию g, для которой справедливо соотношение

$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ .

Если g – тождественная функция ($g(\mu) = \mu$, identity link), то только что написанное соотношение превращается в обычную регрессию. Если же g – это логарифм (log link), то получаем то, что называется *логлинейной моделью*:

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

Преимущество использования логлинейной модели заключается в том, что она дает возможность свести изучение сложных взаимодействий между независимыми переменными (т.е. подбор таких произведений х-ов, которые делают адекватной реальности используемую модель; выше мы говорили о важности и трудности решения этой задачи) к поиску коэффициентов линейной зависимости (поскольку логарифм произведения равен сумме логарифмов).

Особую важность имеет т.н. логит-связь (logit link), когда функция g является функцией вида:

$$g(\mu) = \log \frac{\mu}{1 - \mu}$$

Обобщенная линейная модель при использовании такой связи называется *логит-моделью* (logit model). Эта модель играет большую роль в тех случаях, когда Y – дихотомическая переменная. Используя введенные выше обозначения (p – доля единичных значений Y, а q = (1–p) – доля нулевых значений того же признака) можно сказать, что здесь

$$g(\mu) = \log \frac{p}{q}$$

Другими словами, функция g является логарифмом отношения преобладания. Ниже для простоты будем предполагать, что у нас только один признак X. Уравнение вида

$$\log \frac{p(X)}{1 - p(X)} = \alpha + \beta X$$

называется *логистической регрессионной функцией*. Важность ее изучения представляется очевидной (скажем, для приведенного в предыдущих параграфах примера она позволяет выявить причины изменения соотношения читающих и не читающих данную газету).

Не менее очевидной является важность изучения и т.н. *линейной вероятностной модели*

$$p(X) = \alpha + \beta x$$

(применительно к тому же примеру, здесь речь идет об изменении доли читающих газету). Заметим, что, когда независимых переменных много, подобного рода уравнения совпадают с теми, которые обычно связываются с логлинейным анализом (там в качестве значений независимой переменной выступают частоты многомерной таблицы сопряженности).

Описанные модели являются очень полезными для социолога. Для интерпретации полученных с их помощью результатов можно использовать описанные в п. 2.6.4 приемы. Отличие будет состоять в трактовке того, что стоит в левой части найденного регрессионного уравнения. Эта трактовка определяется тем, что было только что сказано нами. Ясно, что использование упомянутых моделей расширяет круг решаемых с помощью НРА задач.

# ПРИЛОЖЕНИЯ К ЧАСТИ II

## Приложение I

### Разные способы расчета медианы и предполагаемые ими модели

Опишем разные способы расчета медианы на примере.

Предположим, что для 10 школьников значения коэффициента IQ, определенные с помощью шкалы интеллекта Стенфорда-Бине, оказались равными:

113, 120, 119, 115, 122, 126, 120, 112, 120, 119.

Известно, что значением коэффициента может быть любое целое число от 0 до 150. Покажем, каким способами можно рассчитать медиану этого распределения.

Прежде всего необходимо определить тип используемой шкалы. Учитывая, что множество шкальных значений велико и что пороги различимости различий между соседними шкальными значениями для человека (и для респондента, и для социолога) достаточно велики, будем считать, что равенства типа 128-127=113-112 отражают реальность. Поэтому будем считать шкалу интервальной (полагаем очевидным то, что отношения равенства и порядка между шкальными значениями тоже отражают одноименные эмпирические отношения).

Способ расчета медианы и, как следствие, получаемое значение искомой величины определяется модельными соображениями, интерпретацией исходных данных (связанной в первую очередь с нашими представлениями о порождении данных и о соотнесении выборки и генеральной совокупности). Рассмотрим возможные варианты.

а) Выборка – это и есть генеральная совокупность. Кроме названных чисел у нас в принципе ничего нет. Тогда медиану целесообразно найти с помощью вариационного ряда:

112, 113, 115, 119, 119, 120, 120, 120, 122, 126

$$Me = 119{,}5$$

В таком случае естественной будет следующая функция распределения.

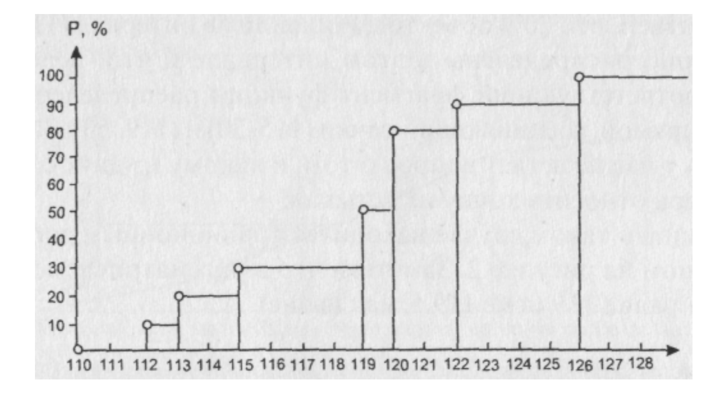


**Рис. 1.** Вид функции распределения при отождествлении выборки с генеральной совокупности

Однако более отвечающей реальности (хотя и опирающейся на непроверяемые модельные соображения) представляется другая функция распределения. В ее основе лежат два предположения. Первое состоит в том. что, вообще говоря, в качестве значения нашей переменной может служит любое действительное число из рассматриваемого диапазона. Подчеркнем, что здесь фактически две посылки: первая состоит в том, что в принципе нам могут встретиться любые целочисленные значения; против нее вряд ли кто-либо будет возражать; вторая же – говорит о возможности встретить нецелочисленные значения. Последняя посылка обычно по вполне понятным причинам вызывает сомнения. Принять ее – значит полагать, что в принципе измеряемая переменная непрерывна, что к ее дискретности приводит несовершенство используемого способа измерения и отсутствие более адекватных измерительных алгоритмов. После принятия указанного предположения функцию распределения естественно представлять следующим образом (отрезки построенной ломаной линии соединяют левые концы стрелок с предыдущего рисунка).

Второе предположение есть предположение о постепенности, равномерности накопления объектов в каждом заданном выборкой интервале. Так, если в процессе построения графика накопленных частот (выборочного аналога функции распределения) в точке Х = 115 у нас "накопилось" 30% объектов, а в точке 119 – уже 50%, то мы считаем, что 20% объектов, попавших в интервал (115, 119), равномерно распределены в этом интервале и что, вследствие этого, соответствующий фрагмент функции распределения есть отрезок прямой, соединяющий точки (115, 30) и (119, 50). Заметим, что здесь у нас не встает вопрос о том, к какому из двух соседних интервалов относить точку их "стыка".

Медиана в таком случае находится традиционным способом, отраженном на рисунке. Заметим, что в рассматриваемой ситуации она равна 119 (а не 119,5, как выше).

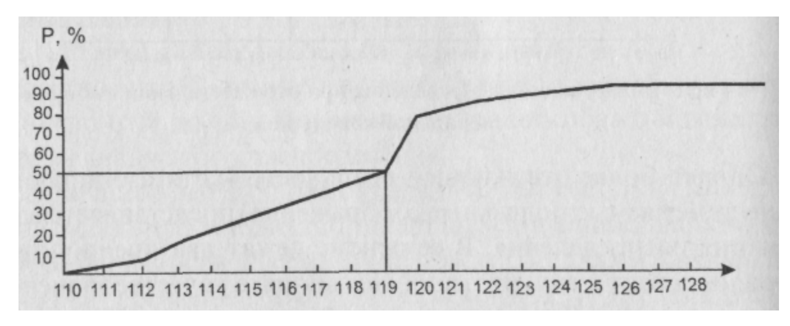


**Рис. 2.** Вид функции распределения при предположениях (а) о непрерывности рассматриваемой случайной величины и (б) равномерном накоплении единиц совокупности в каждом заданном выборкой интервале. Me = 119

На деле социолог обычно пользуется еще более сильным предположением. А именно, при высказанных выше предположениях он задает некоторое разбиение диапазона изменения рассматриваемого признака на интервалы (о встающих здесь проблемах мы говорили в п. 1.1.2) и полагает, что в действительности для него при рассмотрении какого-либо конкретного объекта имеет смысл не то, какое именно значение признака этому объекту отвечает, а то, в какой интервал это значение попадает. При построении выборочного представления функции распределения доля объектов, отвечающих какому-либо интервалу, откладывается, вообще говоря, от любой точки последнего. На следующих двух рисунках отражены наиболее распространенные варианты: на первом – указанная доля откладывается от середины интервала, на втором – от его правого конца. Значения медиан обозначены на рисунках.

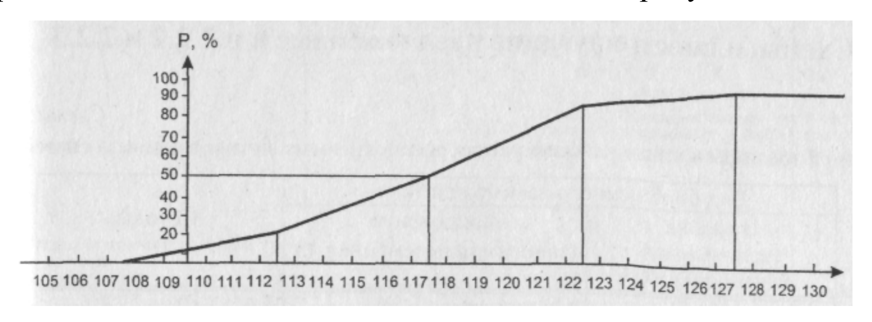



**Рис. 3.** Вид функции распределения при предположениях (а) о непрерывности рассматриваемой случайной величины и (б) заданном априори разбиении на интервалы диапазона ее изменения; (в) отнесении точки “стыка” двух интервалов направо; (г) равномерном накоплении единиц совокупности в промежутке от середины одного интервала до середины другого. Me = 117,5.

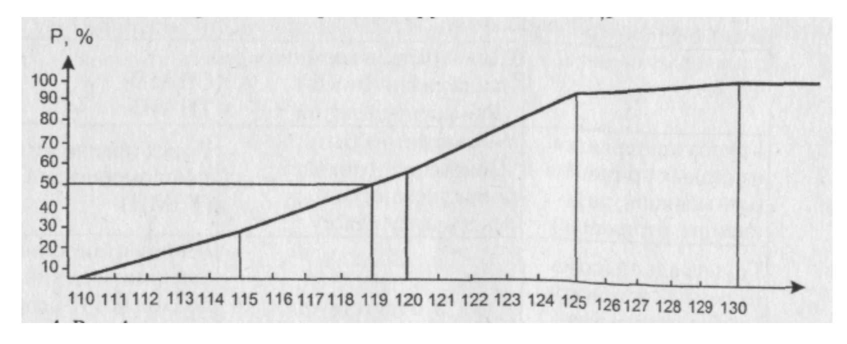



**Рис. 4.** Вид функции распределения при предположениях (а) о непрерывности рассматриваемой случайной величины и (б) заданном априори разбиении на интервалы

диапазона ее изменения; (в) отнесении точки “стыка” двух интервалов направо; (г) равномерном накоплении единиц совокупности в каждом интервале. Me = 119

**Приложение 2**

**Схемы, иллюстрирующие предложенные в п. 2.2.2 и 2.2.3**

*Схема 1.*

**Использованная в книге классификация рассмотренных методов анализа связей**

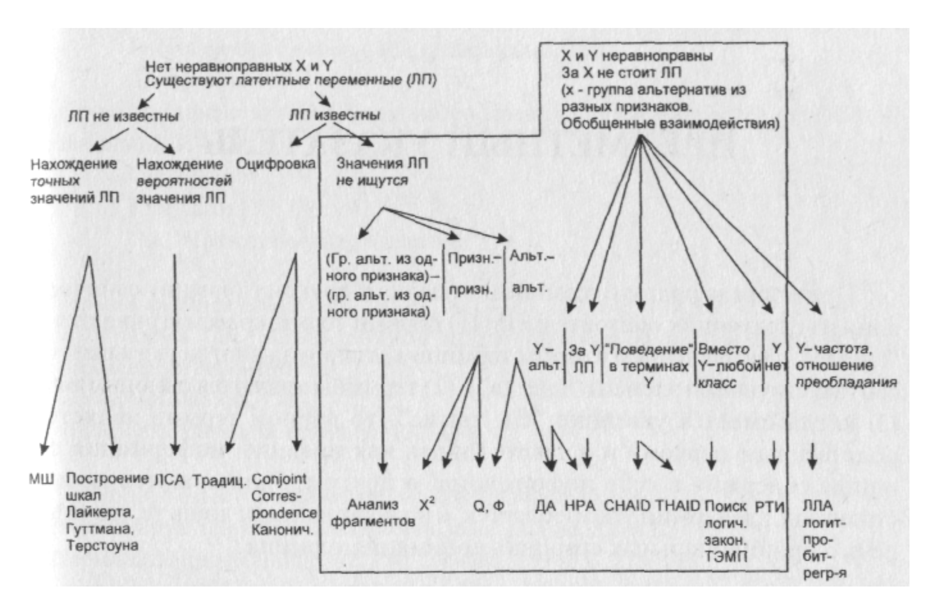
| Вид обобщенного взаимодействия | | Методы |
|---|---|---|
| Посылка (независимая переменная, X) | Заключение (зависимая переменная Y) | |
| Альтернатива | Альтернатива | ДА, Q, Ф |
| Группа альтернатив из одного признака (конъюнкция) | Группа альтернатив из одного признака (конъюнкция) | Анализ фрагментов таблицы сопряженности |
| Группа альтернатив из разных признаков (конъюнкция) | Альтернатива | ДА, НРА с номинальным Y |
| То же | "Поведение" в терминах Y:<br>- сила связи X-ов с Y, CHAID<br>- вид распределения Y THAID | |
| Группа альтернатив из разных признаков (конъюнкция, дизъюнкция отрицание) | Y-ка может не быть. "Поведение" означает принадлежность к некоторому классу | Поиск логических закономерностей (ТЭМП) |
| Группа альтернатив из разных признаков (любая логическая функция) | Y отсутствует | Репрезентационно-аксиоматический подход (РТИ-репрезентационая теория измерений) |
| Один X как целое | Один Y как целое | $\chi^2$, $\lambda$, Q, Ф |

| | | |
|---|---|---|
| Группа X | То же | НРА |

*Схема 2.*

**Классификация рассмотренных методов на базе предположений о существовании латентных переменных.**

(Рамкой обведено то, что рассматривается в учебнике)




Сокращения: ЛП – латентная переменная, гр. альт. – группа альтернатив, МШ – многомерное шкалирование, ЛСА – латентно-структурный анализ, ДА – детерминационный анализ, НРА – номинальный регрессионный анализ, РТИ – репрезентационная теория измерений, ЛЛА – логлинейный анализ.

# ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Приводимая рядом с термином отсылка к другому термину означает одно из следующих обстоятельств: (1) первый термин рассматривается в “гнезде”, озаглавленном вторым термином (в скобках иногда указывается соответствующий элемент “гнезда”); (2) термины являются синонимами; (3) когда имеется указание “см. также”, то второй термин является родственным первому и в тексте книги, как правило, информация об одном содержит в себе информацию о другом. Указываются не все страницы, где термин употребляется, а по возможности лишь те, где идет речь о принципиальных сторонах понимания термина.

## *Литература*


*Адамов С.Ю.* Система анализа нечисловой информации “САНИ” // Социология: 4М (методология, методика, математическое моделирование). 1991. 2. С.86-104

*Айвазян С.А., Мешалкин Л.Д., Енюков И.С.* Прикладная статистика. Т.1.М.: Финансы и статистика, 1983.

Алгоритмы и программы восстановления зависимостей. М.: Наука, 1984

*Аптон Г.* Анализ таблиц сопряженности. М.: Финансы и статистика, 1982 (Upton G.J.G. The analysis of cross-tabulated data. N.-Y.: J.Wiley&Sons, 1978)

*Аргунова К.Д.* Качественный регрессионный анализ в социологии. М.: ИСАН СССР, 1990

*Бартоломью Д.* Стохастические модели социальных процессов. М.: Финансы и статистика,1985

*Батыгин Г.С.* Соотношение понятий и переменных в социологическом сследовании // Социс, 1981, № 3. С. 53-63

*Батыгин Г.С.* Обоснование научного вывода в прикладной социологии. М.: Наука, 1986

*Батыгин Г.С.* Ремесло Пауля Лазарсфельда (Введение в научную биографию) // Вестник АН СССР, 1990, № 8

*Батыгин Г.С., Девятко И.Ф.* Миф о качественной социологии // Социологический журнал, 1994, № 2. С. 28-42

*Божков О.Б.* Письмо в редакцию журнала “Социологические исследования” // Социс, 1988, № 3. С. 135-137.

*Бочаров В.А., Маркин В.И.* Основы логики. М.: Космополис, 1994

*Браверман Э.М., Киселева Н.Е., Мучник И.Б., Новиков С.Г.* Лингвистический подход к задаче обработки больших массивов информации // Автоматика и телемеханика, 1974, №11. С. 73-88

*Бранский В.П.* Теоретические основания социальной синергетики // Петербургская социология, 1997, №3

*Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979

*Витяев Е.Е.* Семантический подход к созданию баз знаний. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных // Логика и семантическое программирование (Вычислительные системы, вып. 146). Новосибирск, 1992

*Витяев Е.Е., Логвиненко А.Д.* Обнаружение законов на эмпирических системах и тестирование систем аксиом теории измерений // Социология: 4М (методология, методы, математическое моделирование), 1998, №10. С. 97-121

*Витяев Е.Е., Москвитин А.А.* ЛАДА – программная система логического анализа данных // Методы анализа данных (Вычислительные системы, вып. 111). Новосибирск, 1985. С. 38-58

*Витяев Е.Е., Москвитин А.А.* Введение в теорию открытий. Программная система DISCOVERY // Логические методы в информатике (Вычислительные системы, вып. 148). Новосибирск, 1993. С. 117-163

*Войшвилло Е.К.* Понятие. М.: Изд-во МГУ, 1989.

*Волошинов А.В.* Пифагор. Союз истины, добра и красоты. М.: Просвещение, 1993.

*Гласс Дж.,Стэнли Дж.* Статистические методы в педагогике и психологии. М.: Прогресс, 1976

*Гмурман В.Е.* Теория вероятностей и математическая статистика. М.: Высшая школа, 1998а

*Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. М.:Высшая школа, 1998б

*Гнеденко Б.В.* Курс теории вероятностей. М.: Наука, 1965

*Голод С.И.* Современная семья: плюрализм моделей // Социологический журнал, 1996, №3/4. С. 99-198.

*Гумилев Л.Н.* Древняя Русь и Великая степь. М.: Мысль, 1993

*Давыдов Ю.Н.* Ближайшие предшественники О.Конта // История теоретической социологии. Т.1. М.: Наука, 1995. С. 190 – 257

*Давыдов Ю.Н.* Идиографический метод // Справочное пособие по истории немарксистской западной социологии. М.: Наука, 1986. С.118-122

*Давыдов Ю.Н.* Н.Д.Кондратьев и вероятностно-статистическая философия социальных наук // Кондратьев Н.Д. Основные проблемы экономической статики и динамики. М., 1991. С.453-523.

ДА-система (Детерминационный анализ). М.: Фирма "Контекст", 1989-1997

*Девятко И.Ф.* Методы социологического исследования. Учебное пособие для вузов. Екатеринбург, изд-во Уральского университета, 1998

*Девятко И.Ф.* Модели объяснения и логика социологического исследования. М.: Институт социологического образования и др., 1996

*Джини К.* Средние величины. М.: Статистика, 1970

*Дидэ Э.* и др. Методы анализа данных. М.: Финансы и статистика, 1985 (Diday E. et collaborateurs. Optimisation en classification automatique. Paris: Institut national de rechercher en informatique et en automatique, 1979)

*Дэвид Г.* Метод парных сравнений. М.: Статистика: 1978.

*Дэйвисон М*. Многомерное шкалирование. М.: Финансы и статистика, 1988.

*Евин И.А., Петров В.М.* О некоторых инвариантах в социологическом моделировании (синергетический подход) // Демократические институты в СССР: проблемы и методология исследований. М., 1991.

*Елисеева И.И.* Статистические методы измерения связей. Л.: ЛГУ, 1982

*Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. М.: Статистика, 1977

*Елисеева И.И., Рукавишников В.О.* Логика прикладного статистического анализа. М.: Финансы и статистика, 1982

*Ермаков С.М., Михайлов Г.А.* Статистическое моделирование. М.: Наука, 1982

*Жамбю М.* Иерархический кластер-анализ и соответствия. М.: Финансы и статистика, 1988 (Jambu M. Classification automatique pour l’analyse des donnees. Paris: Borda, 1978)

*Жмудь Л.Я.* Наука, философия и религия в раннем пифагореизме. С.-Пб.: ВГК, Алетейя, 1994.

*Загоруйко Н.Г.* Эмпирическое предсказание. Новосибирск: Наука, 1979

*Задорин И.В.* Экспертный сценарно-прогностический мониторинг: методологические основания и организационная схема // Вопросы социологии. - 1994. - Вып. 5. С. 27-49.

*Ивченко Г.И., Медведев Ю.И.* Математическая статистика: учебное пособие для ВТУЗов. М.: Высшая школа, 1992

Интерпретация и анализ данных в социологических исследованиях.М.: Наука, 1987

*Калинина В.Н., Панкин В.Ф.* Математическая статистика. М.: Высшая школа, 1998

*Капица С.П., Курдюмов С.П., Малинецкций Г.Г.* Синергетика и прогнозы будущего. М.: Наука, 1997

*Кендалл М.Дж., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973

*Клигер С.А., Косолапов М.С., Толстова Ю.Н.* Шкалирование при сборе и анализе социологической информации. М.: Наука, 1978

*Клишина Ю.Н.* Применение анализа соответствий в обработке нечисловой информации // Социология : 4М (методология, методы, математические модели). 1991,2. С. 105-118

*Клюшина Н.А.* Причины, вызывающие отказ от ответа // Социс, 1, 1990. С. 98-105

*Князева Е.Н., Курдюмов С.П.* Законы эволюции и самоорганизации сложных систем. М.: Наука, 1994

*Ковалев Е.М., Штейнберг И.Е.* Качественные методы в полевых социологических исследованиях. М.: Логос, 1999

*Колемаев В.А., Калинина В.Н.* Теория вероятностей и математическая статистика, М.: Инфра-М,1997

Компьютерное моделирование социально-политических проблем. М.: Интерпракс, 1994.

*Конт О.* Дух позитивной философии // Западно-европейская социология XIX века. М.: МУБиУ, 1996. С. 7-93.

Краткий очерк истории философии. М.: Издательство социально-экономической литературы, 1960

*Кузнецов В.И.* Понятие и его структуры. Методологический анализ. Киев: Ин-т философии НАН Украины, 1997.

*Кун Т.* Структура научных революций. М.: Прогресс, 1975.

*Курдюмов С.П., Малинецкий Г.Г., Потапов А.Б.* Синергетика – новые направления // Новое в жизни, науке и технике. Сер. “Математика и кибернетика”, 1989, №11.

*Лазарсфельд П.Ф.* Измерение в социологии // Американская социология. М.: Прогресс, 1972.

*Лакатос И.* Фальсификация и методология научно-исследовательских программ. М.: Московский философский фонд "Медиум", 1995.

*Лакутин О.В.* Учёт пропущенных данных // Применение математических методов и ЭВМ в социологических исследованиях. М.: ИСИ АН СССР, 1982. С.86-90

*Лакутин О.В., Толстова Ю.Н.* Принципы построения, оценки качества и сравнения коэффициентов связи номинальных признаков. М.: ИСАН СССР, 1990

*Лакутин О.В., Толстова Ю.Н.* Коэффициенты связи номинальных признаков, основанные на моделях прогноза и понятии энтропии. М.: ИС РосАН, 1992

*Лбов Г.С.* Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981.

*Ливанова Т.Н.* Методическое пособие по использованию программы AID3 системы OSIRIS (анализ взаимодействия или поиск структуры качественных данных). М.: ИСАН СССР, 1990

*Литтл Р.Дж., Рубин Д.Б.* Статистический анализ данных с пропусками. М.: Финансы и статистика, 1991

Логика социологического исследования. М.: Наука, 1985

*Максименко В.С., Паниотто В.И.* Зачем социологу математика. Киев: Радяньска школа, 1988.

Математические методы анализа и интерпретация социологических данных. М.: Наука, 1989.

Математические методы в современной буржуазной социологии. М., 1966.

Методы анализа данных. Подход, основанный на методе динамических сгущений / колл. авторов под рук. Э.Дидэ. М.: Финансы и статистика, 1985

*Мирзоев А.А.* Логлинейный анализ социологической информации // Многомерный анализ социологических данных (методические рекомендации, алгоритмы, описание программ).М.: ИСИ АН СССР, 1981. С. 118-131

*Мирзоев А.А.* Применение логлинейного анализа для обработки данных социологических исследований // Математико-статистические методы анализа данных в социологических исследованиях. М.: ИСАН СССР, 1980. С. 49-60

*Миркин Б.Г.* Анализ качественных признаков и структур. М.: Статистика, 1980

*Миркин Б.Г.* Группировки в социально-экономических исследованиях. М.: Финансы и статистика, 1985

Моделирование социальных процессов. М.: Изд-во рос.экон. академии, 1993.

*Монсон П.* Современная западная социология. Теории, традиции, перспективы. С.-Пб.: Нотабене, 1992.

*Мосичев А.В.* Влияние формулировки вопроса на результаты эмпирических социологических исследований (аналитический обзор) // Методология и методы социологических исследований. ИСРосАН, 1996. С. 20-38

*Мостеллер Ф., Тьюки Дж.* Анализ данных и регрессия. М.: Финансы и статистика, 1982

*Никаноров С.П.* Метод концептуального проектирования систем организационного управления // Социология: 4М (методология, методы, математические модели), 1995, №7-8

*Никитина Н.Н.* Философия культуры русского позитивизма начала века. М.: Аспект Пресс: 1996.

*Ноэль Э.* Массовые опросы. Введение в методику демоскопии. М.: Ава-Эстра, 1993.

*Орлов А.И.* Общий взгляд на статистику объектов нечисловой природы // Анализ нечисловой информации в социологических исследованиях. М.: Наука: 1985. С.58-92.

*Орлов А.И.* Асимптотика квантований и выбор числа градаций в социологических анкетах // Математические методы и модели в социологии. М.: ИСИ АН СССР, 1977. С.42-55

*Осипов Г.В., Андрев Э.П.* Методы измерения в социологии. М.: Наука, 1977

*Паниотто В.И., Максименко В.С.* Количественные методы в социологических исследованиях. Киев: Наукова Думка: 1982

*Паповян С.С.* Математические методы в социальной психологии. М.: Наука, 1983

*Пасхавер Б.* Проблема интервалов в группировках // Вестник статистики, 1972, 6

*Патрушев В.Д., Татарова Г.Г., Толстова Ю.Н.* Многомерная типология времяпрепровождения // Социс, 1980, №4. С.133-140

*Петренко В.Ф.* Основы психосемантики. М.: Изд-во МГУ, 1997

*Петренко Е.С., Ярошенко Т.М.* Социально-демографические показатели в социологических исследованиях. М.: Статистика, 1979

*Плотинский Ю.М.* Математическое моделирование динамики социальных процессов. М.: Изд-во МГУ, 1992

*Плотинский Ю.М.* Визуализация информации. М.: изд-во МГУ, 1994

*Плотинский Ю.М.* Теоретические и эмпирические модели социальных процессов. Учебное пособие. М.: Логос, 1998

*По Э.* Рассказы. М.: Художественная литература, 1980

*Поппер К.* Логика и рост научного знания. М.,1983

*Пригожин И.* Философия нестабильности // Вопросы философии, 1991,6. С.46-52

Применение факторного и классификационного анализа для типологизации социальных явлений. Новосибирск: ИЭиОПП СО АН СССР, 1976

Рабочая книга социолога. М.: Наука, 1983

*Ракитов А.И.* Статистическая интерпретация факта и роль статистических методов в построении эмпирического знания. М., 1981

*Ростовцев П.С.* Черно-белый анализ связи переменных // Социология : 4М (методология, методы, математические модели). 1998, №10. С. 73-96

*Ростовцев П.С.* Алгоритмы анализа структуры прямоугольных матриц “пятна” и “полосы” // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 203-214

*Ростовцев П.С.* Черно-белый анализ связи переменных // Анализ и моделирование экономических процессов переходного периода в России. Новосибирск: Иэи ОПП, 1996. С.264-286

*Ростовцев П.С., Костин В.С., Корнюхин Ю.Г., Смирнова Н.Ю.* Анализ структур социологических данных. Устойчивость // Анализ и моделирование экономических процессов переходного периода в России. Новосибирск: ИЭиОПП: 1997.С.174-208.

*Рыбников К.А.* Введение в методологию математики. М.: изд-во МГУ, 1979.

*Сачков Ю.В.* Вероятностная революция в науке (вероятность, случайность, независимость, иерархия). М.: Научный мир, 1999

*Семенова В.В.* Качественные методы: введение в гуманистическую социологию. М.: Добросвет, 1998

*Сиськов В.И*. Об определении величины интервалов при группировках // Вестник статистики, 1971, 12

Социальное исследование: построение и сравнение показателей. М.: Наука: 1978.

Статистические методы анализа информации в социологических исследованиях. М.: Наука, 1979

*Степанов Ю.С.* Понятие // Лингвистический энциклопедический словарь. М.: Сов.энциклопедия, 1990.С. 383-385.

*Степин В.С., Горохов В.Г., Розов М.А.* Философия науки и техники. М.: Контакт-Альфа, 1995

*Суппес П., Зинес Дж.* Основы теории измерений // Психологические измерения. М.: Мир, 1967 (Suppes P., Zinnes J.L. Basic measurement theory // Handbook of mathematical Psychology. V.1. N.Y.- L.: J.Wiley&Sons, 1963. P.1-76)

*Татарова Г.Г.* Методология анализа данных в социологии. М., 1998

*Терборн Г.* Принадлежность к культуре, местоположение в структуре и человеческая деятельность: объяснение в социологии и социальной науке // THESIS. – Т. II. – 1994. – Вып.4

Типология и классификация в социологических исследованиях. М.: Наука, 1982

*Толстова Ю.Н.* Обеспечение однородности исходных данных в процессе применения математических методов // Социс, 1986, №6. С. 149-154

*Толстова Ю.Н.* Математика в социологии: элементарное введение в круг основных понятий (измерение, статистические закономерности, принципы анализа данных). М.:ИСАН СССР, 1990а

*Толстова Ю.Н.* Методология математического анализа данных // Социс, 1990б, №6.С.77-87

*Толстова Ю.Н.* Логика математического анализа социологических данных, М.: Наука, 1991а

*Толстова Ю.Н.* Принципы анализа данных // Социология: 4М (методология, методыа, математические модели), 1991б, №1. С.51-61.

*Толстова Ю.Н.* Анализ социологических данных. М.: ИСРосАН, 1994 (учебная программа)

*Толстова Ю.Н.* Анализ данных // Энциклопедический социологический словарь-справочник. М.: ИСПИ РАН, 1995. С. 18-21

*Толстова Ю.Н.* Модели и методы анализа данных социологического исследования. Учебное пособие. М.: ГАУ им. С.Орджоникидзе, 1996а.

*Толстова Ю.Н.* Роль моделирования в работе социолога: логический аспект // Социология: 4М (методология, методы, математические модели), 1996б, № 7. С. 66-85.

*Толстова Ю.Н.* Обобщенный подход к определению понятия социологического измерения // Методология и методы социологических исследований (итоги работы поисковых проектов 1992-1996 г.г.). М.: ИСоцРАН, 1996в. С. 66-95

*Толстова Ю.Н.* Идеи моделирования, системного анализа "качественной" социологии: возможность стыковки (на примере метода репертуарных решеток) // Социология: 4М (методология, методы, математические модели), 1997, №8. С.66-85.

*Толстова Ю.Н.* Измерение в социологии. М.: Инфра-М, 1998.

*Тюрин Ю.Н.* Непараметрические методы статистики. М.: Знание, 1978

*Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере. М.: Инфра-М, 1998.

*Фёдоров И.В.* Причины пропуска ответа при анкетном опросе // Социс, 1982, 2

*Фелингер А.Ф.* Статистические алгоритмы в социологических исследованиях. Новосибирск: Наука, СО, 1985

Философия и методология науки. – Под ред. В.И.Купцова. М.: Аспект Пресс, 1996

*Хейс Д.* Причинный анализ в статистических исследованиях. М.: Финансы и статистика,1981

*Чесноков С.В.* Детерминационный анализ социально-экономических данных. М.: Наука, 1982

*Чесноков С.В.* Основы гуманитарных измерений. М.: Наука, 1985

*Чесноков С.В.* Основы гуманитарных измерений. М.: ВНИИСИ, 1986

*Штомпка П.* Социология социальных изменений. М.: Аспект Пресс, 1996

*Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988

*Яглом А.М., Яглом И.М.* Вероятность и информация. М.: Гос. Изд-во физ-мат. литературы, 1960

*Ядов В.А.* Стратегия и методы качественного анализа данных // Социология: 4М (методология, методы, математические модели), 1991, №1. С. 14-31.

*Ядов В.А.* Два рассуждения о теоретических предпочтениях // Социологический журнал, 1995, №2. С.70-72.

*Ядов В.А.* Стратегия социологического исследования: описание, объяснение, понимание социальной реальности.М.: Добросвет, 1998.

*Ярская-Смирнова Е.* Социокультурный анализ нетипичности. Саратов: Саратовский технологический университет, 1997.

*Agresti A.* Categorical data analysis. N.-Y.: John Wiley and sons, 1990

*Benzecri J.P.* L' analyse de donnees. Tome 2. L' analyse de correspondences. Dunod, 1973

*Blalock H.M.* Conceptualization and measurement in the social sciences. Beverly hill: Sage, 1982

*Blalock H. M.* Power and conflict: Toward a general theory. Newburg Parc – L. – New Delhi : Sage publ., 1989 (Рецензия М.М.Назарова в: Социс, 1991, № 6. С. 148-150)

*Bluman A.G.* Elementary statistics. W.C.Brown Publishers. 1995

*Clausen S.-E.* Applied correspondence analysis. An introduction. Sage university paper series on Quantitative applications in the social sciences, 07-121. Newbury park, CA: Sage, 1998

*Coleman J.* Foundational of social theory, MA: Harvard University Press, 1990

*Demaris A.* Logit modeling: Practical application. Sage university paper series on quantitative applications in the social sciences, 07-086. Newbury park, CA: Sage, 1992

*Derrick F., Magidson J.* Using CHAID with the gains chart option // Proceedings of the 1992 annual meeting of American stat. Ass., Business and Economics Section, 1992

*Diamantopoulos A., Schlegelmilch D.P.* Taking the fear out of data analysis. The Driden Press, 1997.

*Guttman L.* Measurement as structural theory // Psychometrika, 1971. V.6. Pp. 329-349

*Hardy M.A.* Regression with dummy variables. Sage university paper series on Quantitative applications in the social sciences, 07-093. Newbury park, CA: Sage, 1993

*Hinton P.R.* Statistics Explained. A Guide for social Science Students. N.-J.,L., 1995

*Kachigan S.K.* Statistical analysis. An interdisciplinar introduction to univariate and multivariate methods.N.-Y.: Radius Press, 1986

*Kass G.* An exploratory technique for investigating large quantities of categorical data // Applied Statistics,1980, 29:2, 119-127 (сравнние алгоритмов AID и THAID)

*Kerlinger F.M., Pedhazur E.* Multiple regression in behavioral research. N.-Y., 1973 (см. также: Pedhazur E. Multiple regression in behavioral research: Explanation and prediction. - N.-Y.: Holt, Rinehart and Winston, 1982)

*Krantz D.H., Luce R.D., Suppes P., Tversky A.* Foundation of Measurement. N.Y. - L. : Acad. Press. V.1 - V.3, 1971 - 1990

*Kruscal J.B., Wish M.* Multidimensional scaling. Sage university paper series on Quantitative applications in the social sciences, 07-011. Newbury park, CA: Sage, 1978

*Liebetrau A.M.* Measures of association. Sage university paper series on Quantitative applications in the social sciences, 07-032. Newbury park, CA: Sage, 1989

*Louviere J.J.* Analysing decision making: Metric conjoint analysis. Sage university paper series on quantitative applications in the social sciences, 07-067. Newbury park, CA: Sage, 1988

*Magidson J.* The CHAID approach to segmentation modeling // Handbook of marketing research. Cambridge, Mass.: Blackwell, 1993

*McCutcheon A.L.* Latent class analysis. Sage university paper series on quantitative applications in the social sciences, 07-064. Newbury park, CA: Sage, 1987

*Menard S.* Applied logistic regression analisys. Sage university paper series on Quantitative applications in the social sciences, 07-106. Newbury park, CA: Sage, 1995

*Messenger R.S., Mandell G.M.* A model search technique for predictive nominal scale multivariate analysis // J. Amer. Stat. Ass. 1972. V.67. P.768-773 (алгоритм THAID)

*Morgan J.N., Messenger R.C.* THAID - a sequential analysis program for nominal dependent variables. Ann.Arbor: Institute for social research, 1973

*Neter J., Wasserman W., Kutner M.H.* Applied linear statistical models: regression, analysis of variance and experimental designs. R.D.Irwin inc, 1990.

Questions and answers in attitude survey: experiments on question form, wording and context / Schumann H.,Presser S. Thousand Oaks, Calif.,1996.

*Rudas T.* Odds ratios in the analysis of contingency tables. Sage university paper series on Quantitative applications in the social sciences, 07-119. Newbury park, CA: Sage, 1998

*Sirkin R.M.*. Statistics for the social science. SAGE publ.,1995

*Sonquist J., Morgan J.* Searching for data structure. Ann. Arbor, 1973 (алгоритм AID3)

*Tabachnick B.G., Fidell L.S.* Using multivariate statistics. Harper Collins College Publishers, 1996.

Thompson. Canonical correlation analysis. Sage university paper series on Quantitative applications in the social sciences, 07-047. Newbury park, CA: Sage, 1984

*Walsh A.* Statistical for the social sciences: with computer - based applications. - N.Y.: Harper & Row Publishers, 1990.

*Yandell B.S.* Practical data analysis for the designed experiments. - Texts in statistical science, 1997.

## Наименование некоторых серий западных изданий, содержащих ряд брошюр по анализу данных

Advanced Quantitative techniques in the social sciences

Applied social research methods series

Measurement methods for the social sciences

Qualitative research methods

Quantitative applications in the social sciences

Sage focus editions

Sage library of social research

Sage sourcebooks for the human services series

Texts in statistical science (Chapman & Hall)

### *Учебники по анализу качественных данных*

*Berg B.L.* Qualitative research methods for the social sciences. Boston,1995

*Creswell J.W.* Research design: qualitative and quantitative approaches. Thousand Oaks,Calif.,1994

*Miles M.D., Huberman A.M.* Qualitative data analysis. An expanded Sourcebook. Thousand oaks, London, New Delhi: SAGE Publications, 1994

Questions and answers in attitude survey: experiments on question form, wording and context / Schumann H.,Presser S. Thousand Oaks, Calif.,1996

*Sacks H.* Lectures on conversation. Oxford, Cambridge: Blackwell, 1992

*Schwartz H., Jacobs J.* Qualitative sociology. Free Press, 1979

*Strauss A.L.* Qualitative analysis for social scientists. Cambridge, 1987

*Wolcott H.F.* Transforming qualitative data: Description, analysis, and interpretation, 1994

### *Математические методы в качественной социологии*

Computer-aided qualitative data analysis: theory, methods and practice / Ed. by Kelle U., Prein G., Bird K. - Sage, 1995.

*Pfaffenberger B.* Microcomputer applications in qualitative research. - Qualitative research methods. V. 14, 1988.

*Richards L., Richards T.* The transformation of qualitative method: computational paradigms and research processes. Using computers in qualitative research. - L.: Sage, 1991.

*Tesch R.* Qualitative research.Analysis types &software tools.N.-Y.: The Falmer Press, 1995 (1990)

Using computers in Qualitative research / Ed. by Fielding N.G., Lee R.M. - Sage, 1991

*Walsh A.* Statistical for the social sciences: with computer - based applications. - N.Y.: Harper & Row, Publishers, 1990;

*Weaver A., Atkinson P.* Microcomputing and qualitative data analysis. - Avebury, 1994.

*Weitzman E.A., Miles M.B.* Computer programs for qualitative data analysis. - Sage, 1995.